101 Graphical Techniques

Dr. Kishore K. Das
Dr. Dibyojyoli Bhallarcharjee

Asian Books Private Limited

101

GRAPHICAL TECHNIQUES

Dr. Kishore K. Das

Reader, Deptt of Statistics
Gauhati University, Guwahati

&

Dr. Dibyojyoti Bhattacharjee

Senior Lecturer, Deptt. of Statistics
Gauhati University, Guwahati

Asian Books Private Limited

7/28, Mahavir Lane, Vardan House, Ansari Road,
Darya Ganj, New Delhi-110002

Asian Books Private Limited

Corporate and Editorial office
7/28, Mahavir Lane, Vardan House, Ansari Road, Darya Ganj, New Delhi-110 002.
E-Mail: asian@nda vsnl.net.in; purobi@asianbooksindia.com
World Wide Web: http://www.asianbooksindia.com
Phones: 23287577, 23282098, 23271887, 23259161 Fax: 91-11-23262021

Sales Offices

Bangalore	103, Swiss Complex No. 33, Race Course Road, Bangalore-560 001 Phones: 22200438, Fax: 91-80-22256583, Email: asianblr@airtelbroadband.in
Chennai	Palani Murugan Building No.21, West Cott Road, Royapettah, Chennai-600 014, Phones: 28486927, 28486928, Email: asianmds@vsnl.net
Delhi	7/28, Mahavir Lane, Vardan House, Ansari Road, Darya Ganj, New Delhi-110 002, Phones: 23287577, 23282098, 23271887, 23259161; Fax: 91-11-23262021 E-Mail: asian@nda.vsnl.net.in
Guwahati	6, G.N.B. Road, Panbazar, Guwahati, Assam-781 001 Phones: 0361-2513020, 2635729 Email: asianghy1@sancharnet.in
Hyderabad	3-5-1101/1/B IInd Floor, Opp. Blood Bank, Narayanguda, Hyderabad-500 029 Phones: 24754941, 24750951, Fax: 91-40-24751152 Email: hydasian@hd2.vsnl.net.in, hydasian@eth.net
Kolkata	10 A, Hospital Street, Calcutta-700 072 Phones: 22153040, Fax: 91-33-22159899 Email: calasian@vsnl.com
Mumbai	Showroom: 3 & 4, Ground Floor, Shilpin Centre, 40, G.D. Ambekar Marg, Wadala, Mumbai-400 031; Phones : 22619322, 22623572, Fax: 24159899
Noida	G-20, Sector 18, Atta Market, Noida Ph: 9312234916, Email: asiannoida@asianbooksindia.com
Pune	Shop No. 5-8, G.F. Shaan Brahma Com., Near Ratan Theatre, Budhwar Peth Pune-02; Phones: 24497208, Fax: 91-20-24497207 Email: asianpune@asianbookindia.com

1st Published: 2008
ISBN: 978-81-8412-048-6

Published by **Kamal Jagasia** for Asian Books Pvt. Ltd., 7/28, Mahavir Lane, Vardan House, Ansari Road, Darya Ganj, New Delhi-110 002

Laser Typeset at *P.N Computers Shahdara Delhi-110 032*
Printed at Yash Printograph.

"This page is Intentionally Left Blank"

Dedicated to our beloved Teachers

Dr. Jyoti Prasad Medhi,
Emeritus Professor of Gauhati University (Retired)

Dr. S. B. Nandi,
Professor of Gauhati University (Retired)

Dr. Ashok Kumar Bansal,
Professor of University of Delhi

Dr. Salil Dutta
Professor of Arya Vidyapeeth College (Retired), Guwahati

&

Indu Bhusan Roy
Professor of G.C. College (Retired), Silchar

Preface

The title "101 Graphical Techniques", itself indicates the very purpose of the book. Data visualization or graphical representation of Statistical data is now an emerging subfield of Statistics. Though various graphical techniques existed in statistical literature but some very common graphical techniques are found to replicate themselves in almost all books and journals. But with the advances in the field of technology the use of graphics has increased many times. Today, graphical methods play an important role in all aspects of statistical investigation—it begins with explanatory plots, supports various stages of analysis and helps in the final communication and display of results. These days we see extensive use of graphics in print media, television news, sports coverage, advertisement etc. and this gives us an idea about the development of statistical graphics. Most of the recent versions of statistical software are now producing high-resolution self-explanatory graphics and is adding number of more graphs in every recent version of their product. However, the software packages do not provide sufficient text to support the purpose and interpretation of the relatively less common graphics and this in turn restrict their use. Social scientists can now get their data analyzed by using graphical tools that are handy and easier compared to other statistical methods.

But absence of proper text has hindered the development of graphics and there use is restricted even though software is available for producing the graphs. The graphical techniques, if used can reduce a lot of calculations that is involved with other statistical techniques in reaching to a conclusion. A number of data analyst and research workers are in search of a text which can act as a torch bearer in the world of statistical graphics.

Various universities of Europe and USA have started to develop specialized courses on "Data Visualization" or "Statistical Graphics". The trend is soon going to enter different parts of the globe and this book can be considered a handy material for "Statistical Graphics".

The book comprises of 101 statistical plots that can be used for analysis, display and comparison of data. Data sets are provided with most of the plots and relevant calculation if any are also shown. The variables considered along the axes are highlighted and the interpretation of the graph is also discussed. The uses of each of the graphical tool are forwarded along with some related statistical/graphical tools.

The authors like to put into record the support obtained from the software packages like Statistica, Dataplot, DJB Graphics and MS Excel for drawing the plots. Thanks are also due to the web site http://www.itl.nist.gov/ which encouraged us with the basic idea of writing such a book. We owe indebtness to all the authors whose work provided us necessary inspiration for completing this arduous task.

Thanks are also due to Mr. Debasish Bhattacharjee, Branch Manager, Asian Books Private Limited, Guwahati and our publisher for considering the matter for publication.

Errors which have inadvertently crept in the book are regretted.

Authors

Contents

List of Tables

List of Figures

1. AGE PYRAMID

1.1. Definition and Description

For drawing this diagram we use the first quadrant and second quadrant only. Thus the entire X axis and only the positive part of Y axis is considered. Here the variable taken along Y axis (Age in most cases) is divided into several non-overlapping classes. For each class we can consider two related data sets (say Sex). The number of cases of the data sets under each class is then represented by two horizontal bars one along OX and the other along OX', the length of which are proportional to the value of the obs vations. For example, in case of sex the number of females (or percentage) under each age group is represented along OX' and the number of males (or percentage) under each age group is represented along OX. This plot can be used for comparison for two related (comparable) data sets for various non overlapping classes. The plot can be used for the comparison of the magnitude of each component of a variable for various age groups. Since with the increase of age groups the number of occurrences keeps on decreasing so the length of each of the bars keep on decreasing with increasing age groups, thus taking the shape of pyramid. In case the values in the various age groups are similar for the different sexes then the pyramid symmetrical about the Y axis.

1.2. Working Data

The data used for drawing the Age pyramid is a hypothetical data showing the population of a city classified by age groups for the different sexes. The figures are expressed in thousands.

Table 1.1: Population of a town classified by age and sex

Age Group	*Population in Thousand*	
	Female	*Male*
0 – 10	166.6	178.2
10 – 20	152.3	163.7
20 – 30	195.7	192.2
30 – 40	167.3	167.7
40 – 50	160	161.2
50 – 60	122.3	128.8
60 – 70	122.5	114.3
70 – 80	107.2	69.2
80 – 90	57.2	23.4
90 – 100	9.8	2.3

1.3. Axes

X Axis: This axis is used to represent the values of the response variable. Here we represent the population along the axis. It may be noted that the axes start from zero (0) and extends to $+\infty$ both along OX and OX'.

Y Axis: The independent variable is taken along this axis in most cases it is the Age. The Age is classified into various categories and accordingly the axis is graduated.

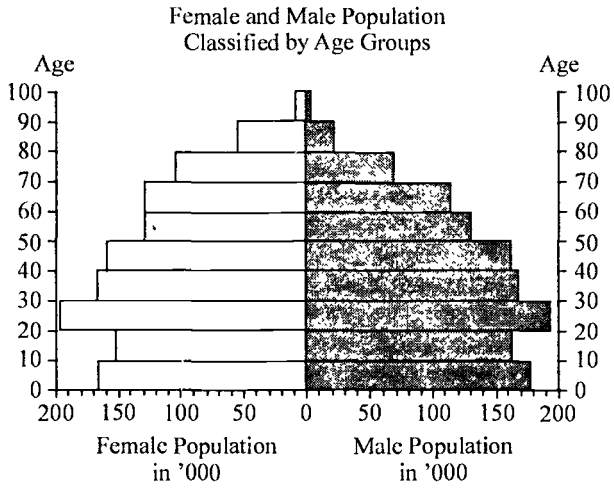


Fig. 1.1. *Age Pyramid representing the data in Table 1.1*

The figure shows that at the higher age group the number of males decreases rapidly compared to their female counterpart. This gives the implication that the males have less longevity than the females.

1.4. Uses

(*a*) It is used to get an idea about the variation of a particular response variable with age for two different categories.

(*b*) The technique can be used to compare the value of response variable with age for different sexes.

(*c*) Sex-wise variation in different demographic variables can be visualized instantly.

1.5. Related Techniques

(*a*) Histogram,

(*b*) Bihistogram,

(*c*) Multiple Bar Diagram, and

(*d*) Grouped Histogram.

2. ANDREWS PLOT

2.1. Definition and Description

The Andrews Plot is an excellent technique of dimension reduction. It was introduced by Andrews (1972). Here a function ($f(\theta)$, say) is used as replacement of a multivariate observation. This is easier to understand compared to other techniques of multivariate graphics as everyone is familiar to plotting of a function $f(\theta)$ with respect to θ.

In an Andrews plot the multivariate observation is first converted into a function, of 'θ' and then is plotted in the graph for various values of θ. Thus, some authors term it as the Andrews Function Plot. Let us consider a p–variate multivariate observation $\mathbf{x}' = (x_1, x_2, \ldots, x_p)$. The Andrews function $f(\theta)$ actually maps the multivariate observation into a function and then displays the function into a two dimensional space. An Andrews curve applies the following transformation to the set of data:

$$f_X(\theta) = x_1/\sqrt{2} + x_2 \sin\theta + x_3 \cos\theta + x_4 \sin 2\theta + x_5 \cos 2\theta + \ldots, \text{ where } -\pi < \theta < \pi \quad \ldots(2.1)$$

The value of θ is made to vary within the specified range with a small increment and accordingly we get a curve for each of the multivariate observation. Thus each multivariate observation is replaced by a curve that runs across the display. The complete display thus consists of several curves (one for each observation) that starts from the extreme left and ends at the other. In the Andrews plot a p dimensional point say $(x_1, x_2, \ldots, x_p)$ gets mapped into a unique trigonometric polynomial. The function is actually a linear combination of the x_i's with changing weights. In the function we find that for every value of θ the linear combination is evaluated with new set of weights for each value of x_i's. For a particular plot we find a one–dimensional projection of all the multivariate observations for a given weight determined by θ. In this plot each curve is a linear combination of all the x_i's of a multivariate observation, where the x_i's are weighted differently by changing the value of θ.

2.2. Working Data

The data in Table 2.1 is about the performance of 30 athletes in Decathlon. The name of the athletes along with their performance in the various events are provided. Here it may be noted that the timed events are measured in seconds and the distances are measured in meters.

Table 2.1: Olympic decathlon data

Name country	*100m*	*Long jump*	*Shot*	*High jump*	*400 m*	*110 m hrd*	*Disc.*	*Pole vlt.*	*Juv*	*1500 m*
Schenk-GDR	11.25	7.43	15.48	2.27	48.9	15.13	49.28	4.7	61.32	268.95
Voss-GDR	10.87	7.45	14.97	1.87	47.71	14.76	44.36	5.1	61.76	273.02
Steen-CAN	11.18	7.44	14.2	1.97	48.29	14.81	43.66	5.2	64.16	263.2

(contd...)

Name country	*100m*	*Long jump*	*Shot*	*High jump*	*400 m*	*110 m hrd*	*Disc.*	*Pole vlt.*	*Jav*	*1500 m*
Thompson-GB	11.62	7.38	15.02	2.03	49.06	14.72	44.8	4.9	64.04	285.11
Blondel-Fra	11.02	7.43	12.92	1.97	47.44	14.4	41.2	5.2	57.46	256.64
Plaziat-FRA	10.83	7.72	13.58	2.12	48.34	14.18	43.06	4.9	52.18	274.07
Bright-USA	11.18	7.05	14.12	2.06	49.34	14.39	41.68	5.7	61.6	291.2
Dewit-HOL	11.05	6.95	15.34	2	48.21	14.36	41.32	4.8	63	265.86
Johnson-USA	11.15	7.12	14.52	2.03	49.15	14.66	42.36	4.9	66.46	269.62
Tarnovetsky -URS	11.23	7.28	15.25	1.97	48.6	14.76	48.02	5.2	59.48	292.24
Keskitalo-FIN	10.94	7.45	15.34	1.97	49.94	14.25	41.86	4.8	66.64	295.89
Gaehwiler-SWI	11.18	7.34	14.48	1.94	49.02	15.11	42.76	4.7	65.84	256.74
Szado-HUN	11.02	7.29	12.92	2.06	48.23	14.94	39.54	5	56.8	257.85
Smith-CAN	10.99	7.37	13.61	1.97	47.83	14.7	43.88	4.3	66.54	268.97
Shirley-AUS	11.03	7.45	14.2	1.97	48.94	15.44	14.66	4.7	64	267.48
Poelman-NZ	11.09	7.08	14.51	2.03	49.89	14.78	43.2	4.9	57.18	268.54
Olander-SWE	11.46	6.75	16.07	2	51.28	16.06	50.66	4.8	72.6	302.42
Freimuth-GDR	11.57	7	16.6	1.94	49.84	15	46.66	4.9	60.2	286.04
Warming-DEN	11.07	7.04	13.41	1.94	47.97	14.96	40.8	4.5	51.5	262.41
Hradan-CZE	10.89	7.07	15.84	1.79	49.68	15.38	45.32	4.9	60.48	277.84
Werthner-AUT	11.52	7.36	13.93	1.94	49.99	15.64	38.82	4.6	67.04	266.42
Gugler-SWI	11.49	7.02	13.8	2.03	50.6	15.22	39.08	4.7	60.92	262 93
Penalver-ESP	11.38	7.08	14.31	2	50.24	14.97	46.34	4.4	55.68	272.68
Kruger-GB	11.3	6.97	13.23	2.15	49.98	15.38	38.72	4.6	54.34	277.84
Lee Fu -TPE	11	7.23	13.15	2.03	49.73	14.96	38.06	4.5	52.82	285.57
Mellado-ESA	11.33	6.83	11.63	2.06	48.37	15.39	37.52	4.6	55.42	270.07
Moser-SWI	11.1	6.98	12.69	1.82	48.62	15.13	38.04	4.7	49.52	261.9
Valenta-CZE	11.51	7.01	14.17	1.94	51.16	15.18	45.84	4.6	56.28	303.17
O'Connel-IRL	11.26	6.9	12.41	1.88	48.24	15.61	38.02	4.4	52.68	272.06
Richards-GB	11.5	7.09	12.54	1.82	49.27	19.56	42.32	4.5	53.5	293.85

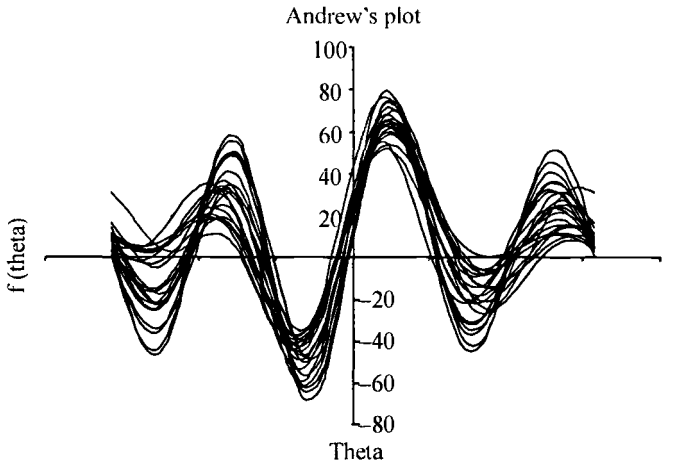


Fig. 2.1. *An Andrews curve for data in Table 2.1*

Andrews curve is order dependent. The first few variables tend to dominate the plot, so it is essential that the most important variables are kept at the beginning. Many authors recommend the use of principal component analysis first in order to determine the relative importance of the variables and then to permute the variables in the curve.

2.3. Axes

X Axis: It can represent the values of θ. Where the value of θ lies between $[-\pi, \pi]$.

Y Axis: The vertical axis is used for representing the function $f_X(\theta)$ given in (2.1).

2.4. Advantages

(*a*) The plot can be used for the detection of multivariate outlier. An observation different from others will be represented by an Andrew's curve which will appear different from all the other curves. It will lie away from the bulk of Andrew's curve.

(*b*) The plot can be used for cluster analysis. The multivariate observations can be divided into clusters using this plot. Similar types of observations will be seen to form similar types of curves and will cluster together.

(*c*) The plot can be used for comparison of several multivariate observations.

2.5. Disadvantages

(*a*) In the plot since all the curves have the same color, so it is difficult to identify which curve represents which observation and thus the use of color may be recommended.

(*b*) In case of the Andrew's plot we perform the linear combination of the variables and hence the values of x_i's ($i = 1, 2, \ldots, p$) are lost and cannot be read from the graph.

2.6. Related Techniques

(*a*) Parallel Co-ordinate Plot, and

(*b*) Icon Plots.

3. ANOM PLOT

3.1. Definition and Description

The chart is similar to the control charts. It is used for the multiple comparisons of the means of several samples. The plot can be drawn for any number of samples greater than 3. ANOM stands for analysis of means. In case the population mean is not known it is estimated and a $(1-\alpha)\times 100\%$ confidence interval of the same is also obtained. The mean of each of the samples are computed and plotted in the graph corresponding sample numbers. The upper limit and lower limit of the confidence interval is plotted in the graph paper in the form of straight lines parallel to X axis. These lines act as confidence bands for the sample means. Now if for a particular sample the mean falls outside the confidence bands then it would imply that the sample mean differs significantly from the others.

3.2. Working Data

The following data is considered originally from Snedecor (1956) which gives the birth weights of Poland China pigs for four liters in pounds.

Table 3.1: Birth weight of poland China pigs in pounds

	Liter 1	*Liter 2*	*Liter 3*	*Liter 4*
	2.0	3.5	3.3	3.2
	2.8	2.8	3.6	3.3
	3.3	3.2	2.6	3.2
	3.2	3 5	3.1	2.9
	4.4	2.3	3.2	2.0
	3.6	2.4	3.3	2.0
	1.9	2.0	2.9	2.1
	3.3	1.6	3.4	—
	2.8	—	3.2	—
	1.1	—	3.2	—
Means	2.84	2.66	3.18	2.67

Overall Mean = 2.8628, Overall Standard Deviation = 0.6894

95% Confidence Interval for mean = [1.51, 4.21].

3.3. Axes

X Axis: The axis is used to represent the sample numbers.

Y Axis: This axis is used to represent the values of the sample mean

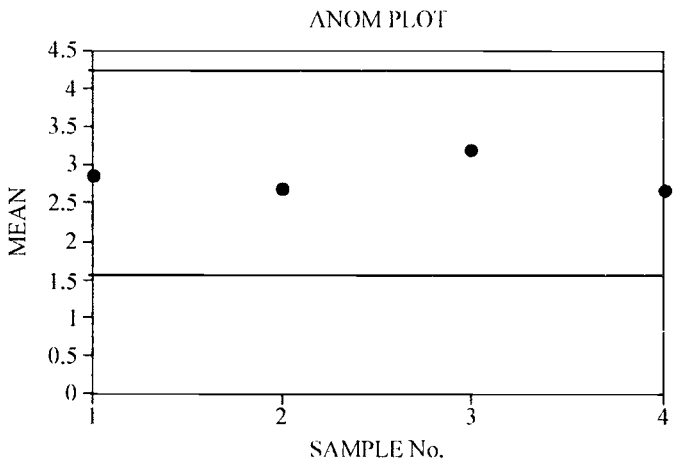


Fig. 3.1. *ANOM plot based on data in Table 3.1.*

From the plot we find that the sample means does not differ significantly at 5% level of significance.

3.4. Uses

(*a*) The plot is simple to draw and easy to interpret.

(*b*) The plot is used for multiple comparison of means from different samples.

(*c*) Even for unequal number of sample sizes this plot can also be used for comparison.

(*d*) The plot can be used as a quick check about the results of ANOVA for one-way classified data.

3.5. Related Techniques

(*a*) Control Chart,

(*b*) ANOVA, and

(*c*) Jittered Plot.

4. AREA CHART

4.1. Definition and Description

The area chart is somewhat similar to a multiple line chart. But here a filled area is plotted in the graph. The area is either bounded by a two lines or by a line and one of the axes. This plot is used for more than one data set. For two response variables, (say the values of the first response variables are plotted and then the plotted points are joined by straight lines or by free hand smooth curve. The values of the second response variable are also plotted in the similar manner and then the area between these two polygons (or curve) is shaded accordingly producing the area chart. This process is repeated for any other response variable if present. In more than one bounded areas the shaded areas are differently colored in order that they can be differentiated.

4.2. Working Data

The data used for this plot is taken from Annual Report, Ministry of Power, Government of India, various issues. The table shows the demand and supply gap in the field of power.

Table 4.1: Demand-Supply gap in the power sector of India

Year	*Requirement*	*Availability*
1990	267632	246560
1991	288974	266432
1992	305266	279266
1993	323252	299494
1994	352260	327281
1995	389721	354045
1996	413490	365900
1997	424505	390330
1998	446584	420235
1999	480430	450594
2000	507213	467401

4.3. Axes

X Axis: The axis is used to represent the independent variable. Here different years are used in this case.

Y Axis: This axis is used to represent the values of the response variable. The variable is amount of electricity in this case.

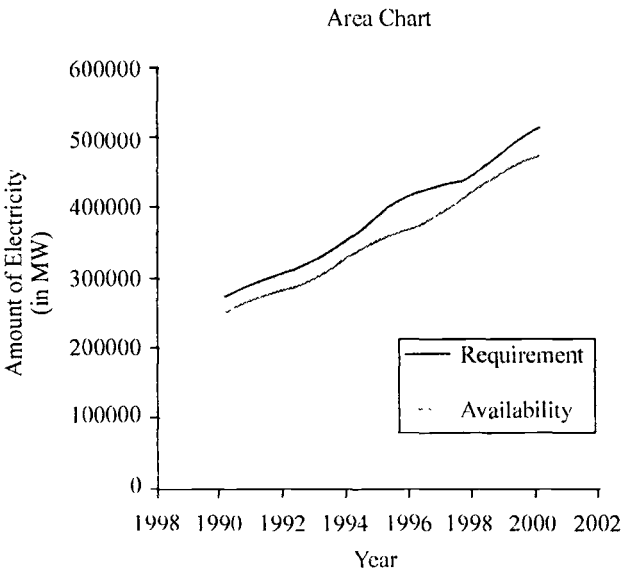


Fig. 4.1. *Area Chart based on data in Table 4.1.*

4.4. Uses

(*a*) The plot provides a visual check related to the consistent variation in the data set.

(*b*) Here the variation in the gap between the two variables can also be studied.

(*c*) The graph can be drawn even without the use of color.

4.5. Related Techniques

(*a*) Line Diagram,

(*b*) Stacked Line Chart,

(*c*) Multiple Line Diagram, and

(*d*) Historigram.

5. ASSOCIATION PLOT

5.1. Definition and Description

This is a plotting technique used for displaying categorical variables in contingency tables. The association plots originally proposed by Cohen (1980) to visualize the cell frequency of a contingency table with the help of rectangles. But before drawing the plot one has to compute the expected values of the cell frequencies of the contingency table and accordingly compute the Pearsonian residuals.

For a two-way contingency table, the expected frequency of each cell, *i.e.*, the frequency that a cell should have under the assumption that the two categorical variables (row and column variables) are independent is computed in the usual manner. If e_{ij} denotes the expected frequency of the cell in the ith row and jth column, then we have

$$e_{ij} = \frac{r_i \times c_j}{n},$$

where r_i = ith row total, c_j = jth column total and n = total frequency.

Let, d_{ij} be the Pearsonian residual corresponding to the ith row and jth column.

So, we have $d_{ij} = \dfrac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}$, where $i = 1, 2, \ldots, r$ and $j = 1, 2, \ldots, c$.

Here, r is the total number of rows and c is the total number of columns.

The value of d_{ij} is obtained for each cell. Now corresponding to each cell a rectangle is drawn. The rectangles in each row are drawn relative to a base line. The positive and negative values of the residuals are differentiated by rectangles growing above or below the base line respectively. The heights of the rectangles are drawn proportional to the corresponding Pearsonian residuals (d_{ij}) and the width are proportional to the square root of expected frequencies ($\sqrt{e_{ij}}$).

We have,

$$\text{length} \propto d_{ij} \text{ and width} \propto \sqrt{e_{ij}}$$

$$i.e., \text{ length} \propto \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}}$$

$$\text{Thus, Area} \propto \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}} \times \sqrt{e_{ij}} = o_{ij} - e_{ij}.$$

Thus the area of each rectangle is proportional to the raw residual. Also, Cohen used different colors for positive and negative residuals. Mayer, Zeileis and Hornik (2003) proposed color and shading pattern similar to those used by Friendly (1994) for the extended mosaic plots can be used in association plot as well. The details of the scheme is available in Plot 69.

5.2. Working Data

The working data is provided in Table 5.1.The data is a two way contingency table derived from a 5-way contingency table used in Edwards and Kreiner (1983). The data come from a sample of employed men aged in between 18 and 67, who were asked whether, in the preceding year, they had carried out any work on their homes.

Table 5.1: A Contingency table showing age of respondent and labour type

	Age		
Labour type	*Less than 30*	*31-45*	*Greater than 46*
Skilled	169	123	68
Unskilled	161	128	160
Office	187	345	250

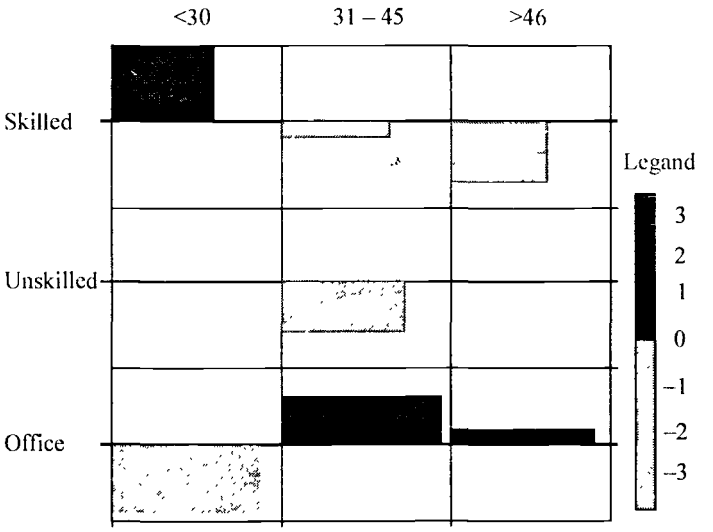


Fig. 5.1. *Cohen's Association Plot based on the data provided in Table 5.1*

5.3. Advantages

(*a*) This plot is useful for visual representation of data arranged in the form of contingency tables.

(*b*) The plot is useful in identification of the cells in which the difference between observed and expected values is more.

5.4. Disadvantages

(*a*) The pattern of shading used in association plot is used to visualize significance but not the pattern of deviation from independence.

(*b*) The plot is not available in most statistical software.

5.5. Related Techniques

(*a*) Chi-square test for independence in contingency table,

(*b*) Mosaic Plot,

(*c*) Sieve Diagram, and

(*d*) Four Fold Display.

6. AUTO CORRELATION PLOT

6.1. Definition and Description

Autocorrelation plot is commonly-used for checking randomness in a set of observations. This plot is also used to check the presence or absence of trend in case of time series data. This randomness is ascertained by computing autocorrelations for data values at varying time lags. If random, such autocorrelations should be near zero for any and all time-lag separations. If non-random, then one or more of the autocorrelations will be significantly non-zero.

Let Y_i ($i = 1, 2, \ldots, N$) be a time series data. The formula for auto correlation of lag h is given by,

$$R_h = \frac{\frac{1}{N}\sum_{i=1}^{N-h}(Y_i - \bar{Y})(Y_{i+h} - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(Y_i - \bar{Y})^2}$$

The values of auto correlation are calculated for various time lags *i.e.*, $h = 1, 2, \ldots$

The values of the various autocorrelation are then plotted against the corresponding lags. In case the data is random then the values of the autocorrelation will lie near to zero, otherwise at least one autocorrelation will be significantly non–zero.

6.2. Working Data

The working data for this purpose is taken from *Newsletter for Pattern Recognition,* 13, No. 3, October 1990. The data gives the number of papers presented in International Conference on Pattern Reorganization, for the years 1981 to 1990 by American authors.

Table 6.1: Table showing papers presented by American authors in International Seminars on Pattern Recognition

Year	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Papers	52	52	81	55	134	63	154	65	62	150

6.3. Axes

X Axis: In this plot along the X–axis we have taken the lags, which are various time lags *i.e.*, $h = 1, 2, \ldots$

Y Axis: The values of the various autocorrelation are considered along the axes. Since autocorrelation lies between -1 to $+1$. So the axis takes the same range as its upper and lower limits.

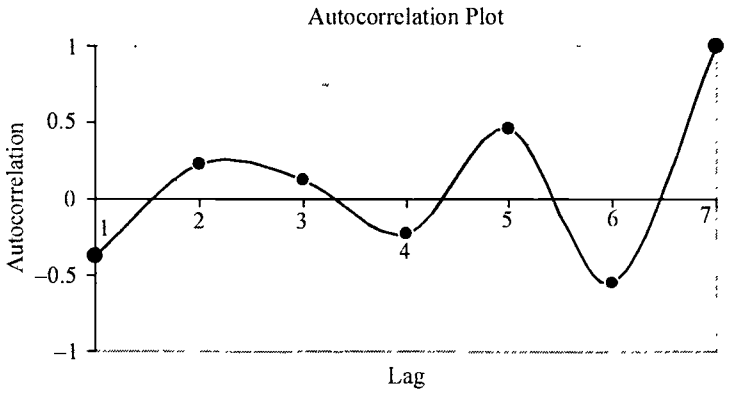


Fig. 6.1. *An Autocorrelation plot for the data in Table 6.1*

6.4. Advantages

(*a*) The main advantage of the plot is to check if the data is random. The randomness is a gate keeper to many statistical tests and techniques. Thus the tool can be considered as a quick check to this feature.

(*b*) It provides a quick check whether the observations are related to an adjacent observation.

(*c*) The autoregressive nature in case of time series data can be put to check in this plot.

6.5. Disadvantages

(*a*) The plot calls for a lot of computation. The computations increases as the number of observations increases in the data set.

(*b*) It may become difficult to identify if the autocorrelation is significantly different from zero.

6.6. Related Techniques

(*a*) Run Test,

(*b*) Run Sequence Plot, and

(*c*) Lag Plot.

7. AUTO COVARIANCE PLOT

7.1. Definition and Description

The plot is used to check if the auto–covariance of samples from the same population changes over different samples. Here the sample identification number is taken along the X axis and the auto covariance of the corresponding sample is plotted along the Y axis. In case the value of the auto–covariance changes to a great extent we find wide variation in the horizontal alignment of the points otherwise the plots remain horizontally aligned.

7.2. Working Data

The data in the table below comprises of 10 samples of size 25 each collected from a uniform distribution [0, 1]. The data is generated MS–Excel.

Table 7.1: Ten samples of size 25 each drawn from uniform [0, 1] distribution

Sam 1	*Sam 2*	*Sam 3*	*Sam. 4*	*Sam 5*	*Sam. 6*	*Sam. 7*	*Sam. 8*	*Sam 9*	*Sam. 10*
0.9831	0 1235	0.2883	0 7567	0.3468	0.2955	0.5258	0.7459	0.0181	0 3976
0.1475	0 9104	0.6808	0.6002	0.8995	0.3238	0.1843	0 6824	0 6781	0 5824
0.6011	0.1696	0.8263	0.9952	0.6893	0 9197	0.1367	0.7565	0.5160	0.8219
0 4804	0.2009	0.7291	0.4706	0.4532	0 6529	0.8259	0.4979	0.8053	0.3093
0 0410	0.6987	0.6024	0.7113	0.2409	0 5726	0.0224	0.1288	0.8210	0.4211
0 6946	0.3501	0.8099	0.6416	0.3980	0.0899	0.5248	0.1656	0.1051	0.6300
0 0421	0.8307	0 1902	0.4848	0 9182	0.1977	0.4556	0.2288	0.3158	0.9681
0 2378	0.3350	0.2684	0.1326	0.3500	0.9334	0 8645	0.8194	0.0712	0.4070
0.7592	0.1647	0.3128	0 1115	0.2569	0.4214	0.0706	0.9864	0.2756	0 5290
0.1494	0.8337	0.1027	0 6087	0.1146	0 0922	0.9811	0.6326	0.1830	0.3182
0 2665	0.7896	0.8965	0.5287	0.3195	0 2282	0.8010	0.3139	0.4535	0.2601
0 1321	0.2349	0 8006	0 5331	0.4688	0 9596	0.3291	0 2953	0.3195	0.3664
0 4519	0.7085	0 2592	0 8883	0.0652	0.3776	0 9567	0.8598	0.4691	0.0639
0.4493	0.0880	0.1384	0.5918	0.6148	0.5797	0.7255	0.6852	0.4227	0.1740
0.4011	0.9376	0.1912	0.1839	0.0987	0.3344	0.9504	0 3197	0.0772	0.4056
0 4983	0.1075	0 5485	0.0540	0.4115	0.6550	0.1022	0.3787	0.4990	0.9310
0 4137	0 9754	0 9882	0.9159	0.6110	0.2300	0.8191	0.6285	0.5895	0.6524
0.3747	0.2912	0.4099	0.8729	0.9804	0.3777	0.9170	0.6814	0.8781	0.4911
0 3344	0.7167	0 1322	0.7988	0.2811	0.4214	0 8295	0.1363	0.3364	0.2564
0.1074	0.3058	0.0576	0.6763	0.3759	0 6519	0.8686	0.1907	0.4609	0.6086

(contd...)

Sam. 1	*Sam. 2*	*Sam. 3*	*Sam. 4*	*Sam. 5*	*Sam. 6*	*Sam. 7*	*Sam. 8*	*Sam. 9*	*Sam. 10*
0.3889	0.7236	0.1997	0.7338	0.8543	0.3691	0.6659	0.8036	0.4260	0.2792
0.7984	0.0295	0.7733	0.9838	0.6660	0.4220	0.4366	0.4855	0.9450	0.3947
0.6808	0.5769	0.5336	0.3224	0.6764	0.0230	0.6389	0.1733	0.9159	0.9290
0.5559	0.3601	0.5832	0.6026	0.5462	0.3623	0.8957	0.0745	0.1783	0.7832
0.7694	0.4820	0.2005	0.2894	0.4024	0.2990	0.9856	0.1673	0.5011	0.8253

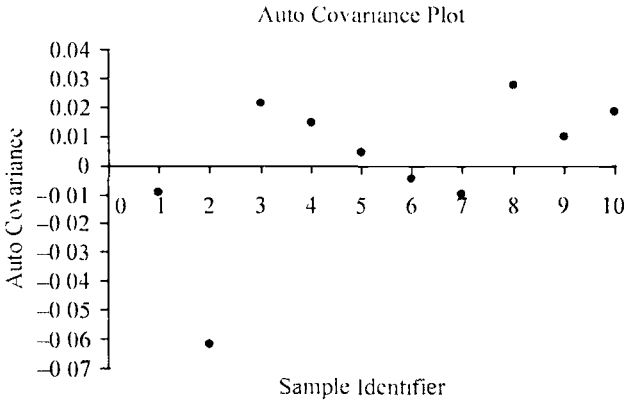


Fig. 7.1. *An Auto Covariance plot for the data in Table 7.1*

The plot shows that the auto covariance's shows wide variation over the samples.

7.3. Axes

X Axis: The axis is used for representing the sample identification number.

Y Axis: The vertical axis is used to represent the auto covariance of the different samples.

7.4. Advantage

(*a*) The main advantage of the plot is to check if the auto-covariance of samples from the same population changes over different samples.

(*b*) It provides a quick check whether the observations are related to an adjacent observation.

7.5. Disadvantage

(*a*) The plot calls for a lot of computation. The computations increases as the number of observations increases in the data set.

(*b*) Statistical software does not in general provide the option of drawing the plot.

7.6. Related Techniques

(*a*) Auto Correlation Plot, and

(*b*) Auto Covariance.

8. BAR CHART (SIMPLE)

8.1. Description of the Plot

The bar chart or bar diagram consists of group of equally spaced rectangular bars, one for each category (or class) fo given statistical data. The bars starting from a common base line, have equal width but their lengths are proporitonal to the values of statistical data that they represent. Thus, it is only the length of the bar that varies with change in the variable and so, the bar diagram is a one-dimensional diagram. The bars are equally spaced from each other along the axis from which it originates. Though the diagram uses both the axes but it is one dimensional as its originating axes is not ordered but is only used space the bars out.

The bars may be placed vertically or horizontally. Generally vertical bars are used to represent time series data or data classified by the values of a variable while horizontal bars are used to depict data classified by attributes only.

8.2. Working Data

The working data consists of come hypothetical values related to employment of individuals in various Public Sectors in the state. The data is in thousands.

Table 8.1: Distribution of persons employed in various public sectors in thousand

Public Sectors	*Persons Employed ('000)*
Central Govt.	75
State Govt	123
Semi Govt.	54
Local Bodies	18

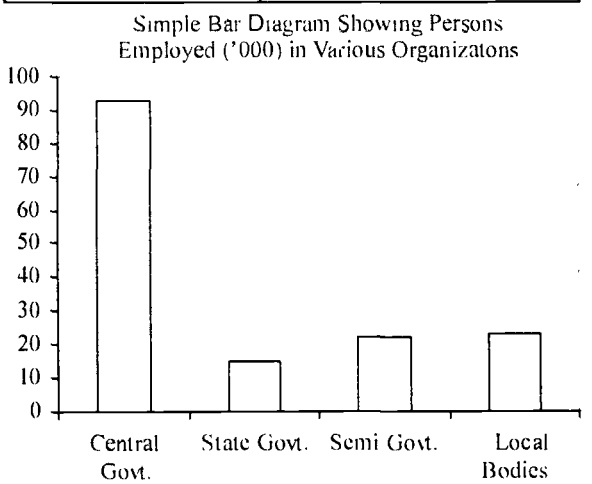


Fig. 8.1. *A simple bar diagram for the data in Table 8.1*

8.3. Axes

X Axis: It represents the various categories corresponding to which the values are provided.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took the number of persons along the Y axis.

The bars may be allowed to extend from the Y axis as well and in such a case the representation of the axes will get inverted.

8.4. Uses

(*a*) Bar diagrams are widely used techniques of data representation especially in cases where the magnitudes of one or more variables under different categories (or classes) are to be compared.

(*b*) They are simple to draw and almost all the statistical software provide the option of drawing it.

8.5. Related Techniques

(*a*) Pie Diagram,

(*b*) Multiple Bar Diagram,

(*c*) Sub-divided Bar Diagram,

(*d*) Histogram, and

(*e*) Impulse Chart.

9. BAR CHART (MULTIPLE)

9.1. Description of the Plot

The multiple bar chart or diagram is developed from simple bar diagram (See Plot 9). Here we have two or more sets of numerical information under various categories. Thus when the data has several categories where each categories have a multiple number of sub–categories common in each of these categories, then we can use the multiple bar diagram. The categories and sub–categories are taken along the X axis. Here, consecutive bars belonging to various sub–categories remains attached to each other. The bars start from the same base *i.e.*, X axis and proceed along the Y axis, with its length being proportional to the value of the response variable corresponding to each sub–category. However, the widths of the bars are equal. Any two categories are separated by an equal width to understand the difference between the various categories. Bars corresponding to different sub–categories are differently shaded and/or colored.

Sometimes the bars are also placed horizontally and not vertically. Generally vertical bars are used to represent time series data or data classified by the values of a variable while horizontal bars are used to depict data classified by attributes only.

9.2. Working Data

The working data for this diagram is the outlay of expenditure during the second and third five year plans of India is several sectors.

Table 8.1: Data showing the outlay of expenditure during the second and third five year plans

Heads	*Total Expenditure (in Rs Crores)*	
	Second Plan	*Third Plan*
1. Agriculture and community development	530	1068
2. Irrigation and power	865	1662
3. Villages and small industries	175	264
4 Organised industries and minerals	900	1520
5. Transport and communication	1300	1486
6. Social service and miscellaneous	830	1500

9.3. Axes

X Axis: It represents the various categories and sub–categories corresponding to which the values are provided.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took Rupees (in Crores) along the Y axis.

The bars may be allowed to extend from the Y axis as well and in such a case the representation of the axes will get inverted.

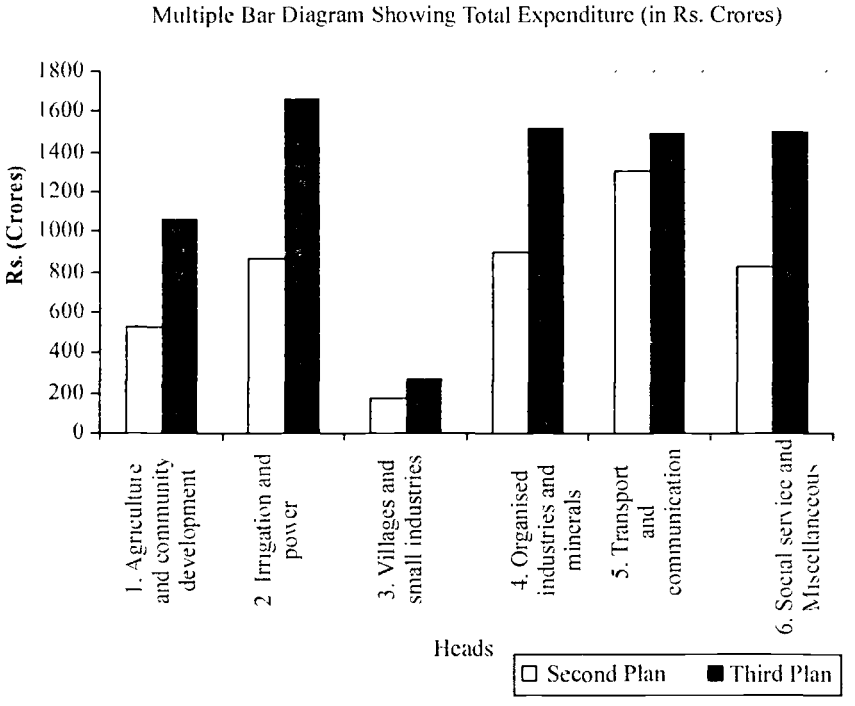


Fig. 9.1. *A multiple bar diagram for the data in Table 8.1*

9.4. Uses

(*a*) This type of bar diagrams is widely used techniques of data representation especially in cases where the magnitudes of one or more variables under different categories and sub-categories are to be compared.

(*b*) They are simple to draw and almost all the statistical software provides the option of drawing it.

9.5. Related Techniques

(*a*) Simple Bar Diagram,

(*b*) Sub-divided Bar Diagram,

(*c*) Histogram, and

(*d*) Impulse Chart.

10. BAR CHART (SUB DIVIDED)

10.1. Description of the Plot

Sub-divided bar chart is also called as sub–divided bar diagram or component bar diagram. Such diagrams are used for comparing the sizes of a given response variable under different categories along with the different component parts among themselves. Thus a relation between each part and the whole may also be visualized. Here based on the value of the response variable under a particular category a thick bar is drawn with its base as X–axis. This bar is divided into several components based on their magnitude in a cumulative fashion. The components are then variously shaded or colored so that the components can be differently identified. The legend accompanying the diagram explains the shade/color used for different components. The procedure is repeated for each and every category. The bars may be placed vertically or horizontally but mostly vertical bars are used.

They are equally spaced from each other along the axis from which it originates. Though the diagram uses both the axes but it is one dimensional as its originating axes is not ordered but is only used to space the bars out. However the bars are difficult to draw in case there are a large number of heads or categories.

10.2. Working Data

The working data consists of the total expenditure by the Indian Government during the First and Second five year plans in the field of education. The components of the total expenditure under various heads are also shown.

Table 10.1: Data showing the total outlay of expenditure during the first and second five year plans

Heads	*Total Expenditure (in Rs Crores)*	
	First Plan	*Second Plan*
1. Elementary Education	93	89
2. University Education	15	57
3. Secondary Education	22	51
4. Technical Education	23	48
5. Social Education	5	5
6. Administrative and Miscellaneous	11	57

10.3. Axes

X Axis: It represents the various categories or heads corresponding to which the values are provided.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took Rupees (in Crores) along the Y axis.

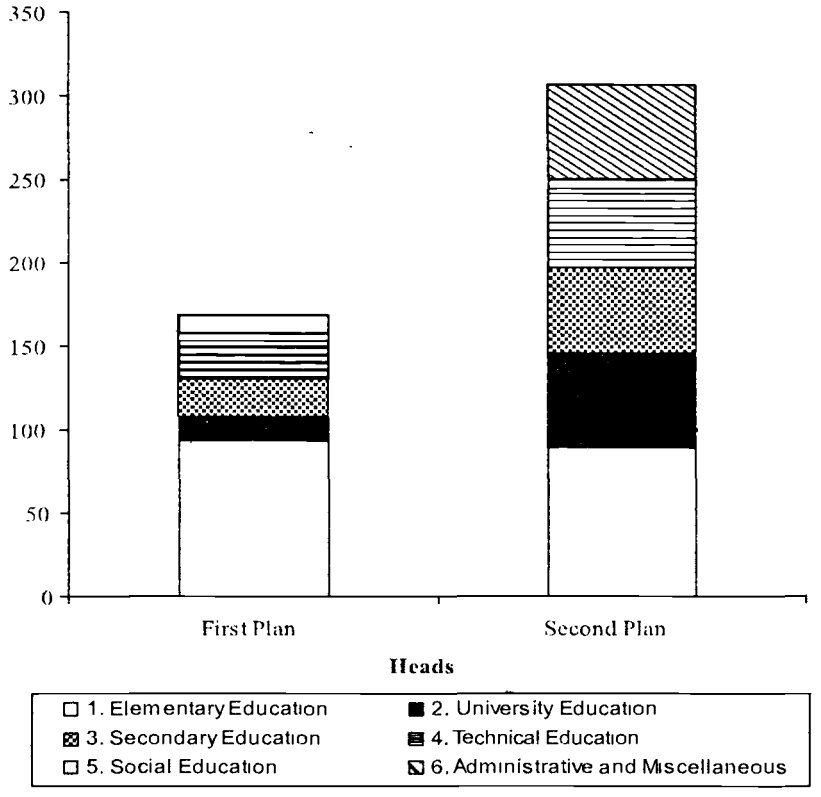


Fig. 10.1. *Sub-divided Bar Diagram for the data in Table 10.1*

10.4. Uses

(*a*) This type of bar diagrams is widely used techniques of data representation especially in cases when the values of the response variable are to be compared along with the component parts.

(*b*) This diagram is simple to draw and almost all the statistical software provides the option of drawing it.

10.5. Related Techniques

(*a*) Simple Bar Diagram,

(*b*) Multiple Bar Diagram,

(*c*) Histogram,

(*d*) Impulse Chart, and

(*e*) Pie Chart.

11. BIHISTOGRAM

11.1. Definition and Description

This is an important graphic tool, which mainly helps in the comparison of two data sets for location and dispersion. This can be used as a graphical alternative to *t*-test (for location) and *F*-test (for dispersion). In addition to this, the test can be used for the detection of outliers. The plot enables us to understand the distributional features like skewness and kurtosis of the two data sets that are plotted. This plot can be used in designs of experiment for assessing the factors that has two levels. The plot can be of special importance in medical science to visualize the before versus after characteristics of a treatment. The two histograms when viewed separately gives the viewer an idea about probability models (normal, log-normal, etc.) for individual data sets, and to detect unexpected and unusual values in the data.

The plot consists of two histograms, one above the X-axis and one below the X-axis. The histogram above the horizontal axis is the histogram of the first variable and the one below the horizontal axis is the histogram of the second variable. A variation of the bihistogram is the relative bihistogram, where relative frequencies of both the data sets are plotted.

11.2. Working Data

The sample data for the plot is taken from Anderson (1958). The data gives the measurement (in mm) of head length of the first two adult sons in a few families.

Table 11.1: Head length of first and second sons of some families

Head length of first son(in mm):	191, 195, 181, 183, 176, 208, 189, 197, 188, 192, 179, 183, 174, 190, 188, 163, 195, 196.
Head length of second son(in mm).	179, 201, 185, 188, 171, 192, 190, 189, 197, 187, 186, 174, 185, 195, 187, 161, 173, 185.

11.3. Axes

Above X-Axis: Histogram of the response variable for the first data series. For this data series we consider the head length of the first son.

Below X-Axis: Histogram of the response variable for the second data series. For this data series we consider the head length of the second son.

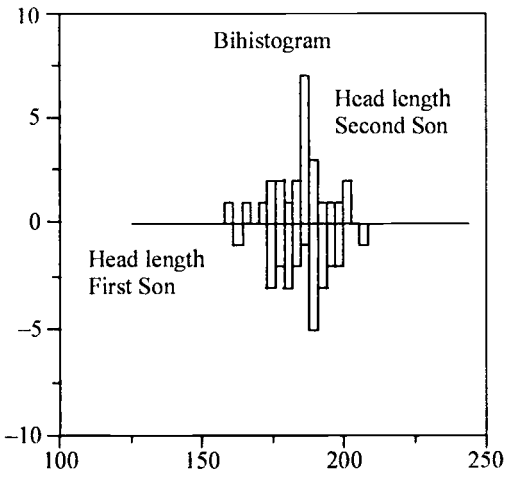


Fig. 11.1. *A bihistogram for the data in Anderson (1958)*

Thus from the bihistogram we find that there is not much variability for the histograms in case of central values and dispersions. But peakedness appears to be more in case of Head length of second son than the first. It implies that the mode of the second series is more compared to the first. Both the series appears to be a little left skewed.

11.4. Advantages

(*a*) The plot can be used for checking simultaneously several tests like- shift in location, shift in dispersion, change in symmetry/skewness, outliers, comparison of mode etc.

(*b*) The plot is based on histograms, which is simple, this makes the plot easily understandable.

(*c*) The plot helps us to study the change in the probabilistic model (if any) of the two data sets in the before versus after cases.

11.5. Disadvantages

(*a*) It is restricted to assessing factors that have only two levels.

(*b*) It's use is restricted as most statistical software does not provide the facility to draw this plot.

11.6. Related Techniques

(*a*) Block Plot,

(*b*) ANOVA,

(*c*) *t*-test, and

(*d*) *F* test.

12. BINOMIALNESS PLOT

12.1. Description of the Plot

Hoaglin (1980) suggested a plot for checking the goodness of fit of a binomial distribution. Let $X \sim Binomial\ (n, p)$, then we have,

$$P(x) = {}^nC_x (1-p)^{n-x}$$

$$= {}^nC_x \left[\frac{p}{1-p}\right]^x (1-p)^n \qquad ...(12.1)$$

Now, $$\eta_x = N \cdot P(X=x) = N \cdot {}^nC_x \left[\frac{p}{1-p}\right]^x (1-p)^n \qquad ...(12.2)$$

Taking log of both the sides,

$$\log \eta_x = \log N + \log ({}^nC_x) + x \log\left[\frac{p}{1-p}\right] + n \log (1-p)$$

$$\Rightarrow \quad \log \eta_x - \log ({}^nC_x) = \log N + n \log (1-p) + x \log\left[\frac{p}{1-p}\right]$$

$$\Rightarrow \quad \log\left[\frac{\eta_x}{{}^nC_x}\right] = \log N + n \log (1-p) + x \log\left[\frac{p}{1-p}\right] \qquad ...(12.3)$$

which is a linear function in terms of x with $\log\left[\frac{p}{1-p}\right]$ as the slope and '$\log N + n \log (1-p)$' as the intercept. Here we call the function

$$\phi(x) = log\left[\frac{\eta_x}{{}^nC_x}\right] \text{ as count metameter function.}$$

For drawing the binomialness plot, we compute the value of $\log\left[\frac{o_x}{{}^nC_x}\right]$, where o_x denotes the values of the observed frequencies. The points $\left(x, \log\left[\frac{o_x}{{}^nC_x}\right]\right)$ are then plotted in a graph paper along with the points $\left(x, \log\left[\frac{\eta_x}{{}^nC_x}\right]\right)$, which falls in a straight line and accordingly these points are connected to get the straight line. In case the observed

data follows a binomial distribution then the plotted points *i.e.*, $\left(x, \log\left[\frac{o_x}{{}^{n}C_x}\right]\right)$ lies in close proximity to the line.

12.2. Working Data

The data that is used for the plot is taken from David (1971), which gives the number of fishes that are caught in one trap. In order to test if the data is a good fit for the binomial distribution we calculate the expected frequency in the usual procedure and written in the third row of the Table 12.1.

Table 12.1: Fish catch data from david (1971)

Fish per Trap	0	1	2	3	4	5	6	7	8
Frequency	1	2	11	20	29	23	10	3	1
Expected Frequency	1	3	11	22	27	22	11	3	0

12.3. Calculations and Figure

Since the value of the parameter is not provided so one can estimate the value of p, the probability of success, in the usual procedure by using the maximum likelihood estimator which is given by,

$\hat{p} = \frac{\bar{x}}{n}$, where $\bar{x} = \frac{\sum f_x o_x}{\sum f_x}$ and accordingly estimate the expected frequencies.

Table 12.2: Calculations related to goodness of fit test for data from David (1971)

Counts (x)	*Observed Frequency* (o_x)	*Expected Frequency* (η_x)	$\log_e\left(\frac{o_x}{{}^{n}C_x}\right)$	$\log_e\left(\frac{\eta_x}{{}^{n}C_x}\right)$
0	1	1	0	0
1	2	3	–1.38629	–0.98083
2	11	11	–0.93431	–0.93431
3	20	22	–1.02962	–0.93431
4	29	27	–0.8812	–0.95266
5	23	22	–0.88986	–0.93431
6	10	11	–1.02962	–0.93431
7	3	3	–0.98083	–0.98083
8	1	0	0	

12.4. Axes

X Axis: It represents the value of the independent variable. In the figure we take the number of fishes per trap along the axis.

Y Axis: The vertical axis consists of the values of the count metameter function which is given by $\log\left[\frac{\eta_x}{{}^{n}C_x}\right]$.

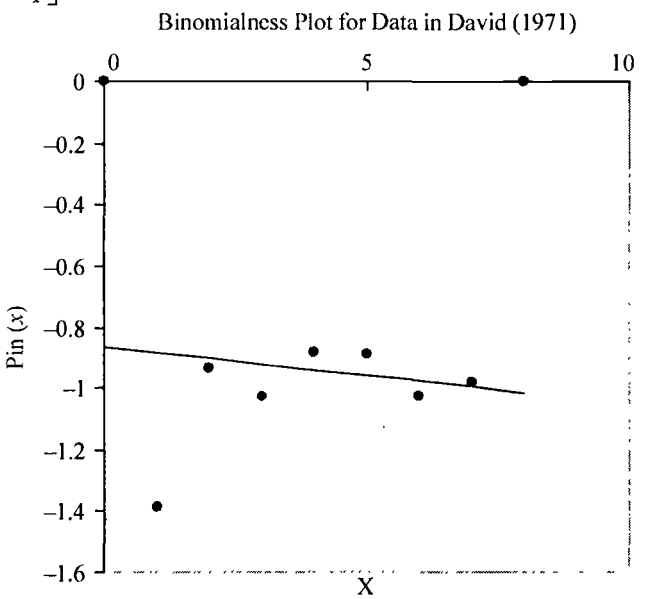


Fig. 12.1. *A binomialness plot for the data in Table 12.1*

Here we find that some of the points fall apart from the hypothetical line though most of the points are close proximity to the line. This indicates that there is an indication of lack of fit to a binomial distribution.

12.5. Uses

(*a*) This plot as the name indicates can be used for checking the goodness of fit of the data to a binomial distribution. The use is restricted and it is difficult to use the plot in deciding about the lack of fit from looking at the binomialness plot. Some confidence bands around the observed line can be a step towards the greater acceptability of the technique.

(*b*) With a little modification the plot can be used to detect the goodness of fit of a data to a truncated binomial distribution.

(*c*) Probably none of the statistical software posses the plot by default.

12.6. Related Techniques

(*a*) Poissonness Plot,

(*b*) Chi-square test for Goodness of Fit, and

(*c*) Probability Plot.

13. BLOCK PLOT

13.1 Definition and Description

The block plot was initiated by Filliben and others (1993). This plot is generally useful if the primary factor has two levels only. The levels are coded with indices. This plot replaces the analysis of variance test by a less assumption dependent binomial test.

The Y axis of the plot consists of the response variable and the X axis consists of all combinations of the secondary factors. The graph comprises of some floating rectangles. The base of the rectangle starts at the lower value of the primary factor and ends at the higher value of the primary factor for a particular combination of the secondary factors. The codes for the various levels of the primary factors (which are generally integers) are written in the blocks in order to identify its position.

Fig. 13.1 is a block plot drawn based on the data given in Table 13.1. The grey vertical lines within the plot are used to partition the different blocks. Each block consists of four rectangles one for each treatment. The values of the treatments at the two levels determine the position and length of the rectangles. The integers '1' and '2' within each rectangle is used to represent the relative position of the levels of the treatments.

13.2. Working Data

The working data is taken from Barlett (1936) which gives the values of number of surviving latherjackets for different controls and emulsions.

Table 13.1: Counts of surviving latherjackets for different controls and emulsions each at two levels

Block \ Treatment		Emulsion 1	Emulsion 2	Emulsion 3	Emulsion 4
1	Level 1	6	17	8	12
	Level 2	10	8	11	17
2	Level 1	4	3	15	6
	Level 2	7	2	20	4
3	Level 1	4	6	10	12
	Level 2	12	3	7	10
4	Level 1	5	1	17	5
	Level 2	5	1	26	8
5	Level 1	2	2	14	12
	Level 2	6	5	11	12
6	Level 1	17	6	22	16
	Level 2	11	5	30	4

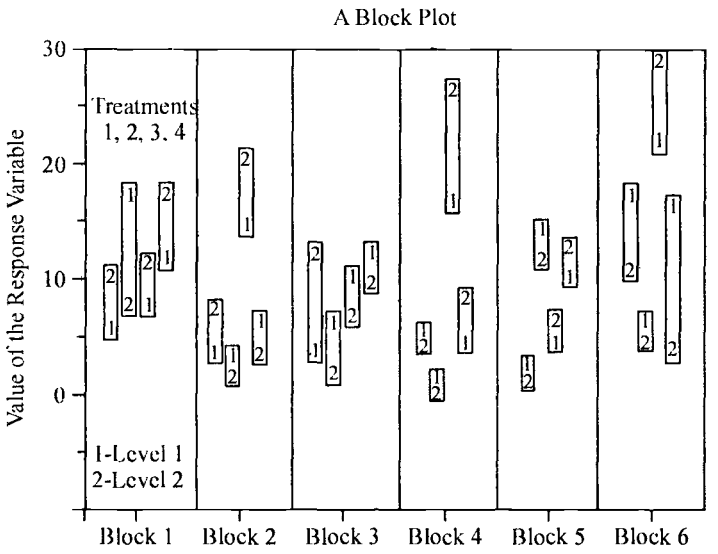


Fig. 13.1. *A Block Plot to the data in Table 13.1*

From the diagram we can find that there are atleast 11 cases out of 24 cases where the performance of Level 1 is better than Label 2. Also from the diagram we can find out that the performance of the various treatments has more variation in the fourth and sixth block. Also the variability is minimum in case of block 3. The third treatment under block 2, 4 and 6 appears to produce high response and so the third treatment can be considered as outliers in some of the blocks.

13.3. Axes

X Axis: It can represent all possible combinations of all factors. For this data set we consider the different blocks along the X axis.

Y Axis: The vertical axis is used for representing the values of the response variable. For this case we took the number of surviving latherjackets along the Y axis.

13.4. Advantages

(*a*) The main purpose of the block plot is to compare different levels classified for different treatments and in the different blocks. Thus we can judge whether there is any significant difference in the response variable for a particular factor under interest called the primary factor and if there is any significant change in the response variable for all other secondary factors (like treatments, blocks).

(*b*) Also these plots can be used for the detection of outliers which is otherwise not possible in ANOVA technique. Once again for a complicated design the graphical representation appears to be much easy for any type of user but the techniques of ANOVA start becoming complex along with the complexity of the design and may even become difficult for people having sufficient mathematical background.

(*c*) ANOVA can be applied only if the data satisfies certain assumptions but for a graphical representation no such assumption is required.

13.5. Disadvantages

(*a*) The block plot is difficult to draw manually.

(*b*) Commonly used statistical software does not provide the option of drawing the plot.

13.6. Related Techniques

(*a*) Jittered Plot,

(*b*) ANOVA, and

(*c*) Bihistogram.

14. BOX PLOT

14.1. Description of the Plot

The box plots are simple to draw and can provide the visualization of at least five values from the set of data. These are median, upper and lower extreme values, upper quartiles and the lower quartiles. The box plots can be of use for producing data summaries and can also be used for data analysis. Here several data sets are plotted and we get a box for each of these data sets. The boxes when viewed in unison acts as an excellent graphical tool for comparing the variation in the data sets and also for the location parameter. The Box plot was introduced by Spear in 1952 and was taken up lately by Tukey (1970, 1977) and was popularized by him.

In order to draw a basic box plot for the representation of a set of observations one may abide by the following steps:

1. To compute the median, the lower quartile and the upper quartile of the data.
2. The values of the variate are taken along the vertical axis and the category or serial number of the data set along the horizontal axis.
3. Plot the median, upper quartile and lower quartile. Draw a box (rectangle) between the upper quartile and lower quartile. The box represents the inter quartile range, that is 50% of the data that lies at the center. A line segment sections the box into two halves represents the position of the 50^{th} percentile i.e. the median.
4. A line is drawn from the lower quartile to the minimum point and another line is drawn from the upper quartile to the maximum point. These two lines are called as the whiskers of the box plot. The box plot is thus also called as the box and whisker plot.

One may draw a box plot with a single box representing one data set only or can draw a box plot with multiple boxes for representing all the data sets taken together.

14.2. Working Data

The data provided in Table 14.1 is self explanatory and the plot below corresponds the Box plot to the data with the years as the categories.

Table 14.1: Index of industrial production: Sector-wise
(Base Year: 1993 – 94 = 100)

Industry	*1997-98*	*1998-99*	*1999-2000*	*2000-01*
Mining	126.4	125.4	126.7	131.4
Manufacturing	142.5	148.8	159.4	167.9
Food Products	133.8	134.7	140.3	154.5
Tobacco	158.1	178.5	192.1	200.4

Industry	*1997-98*	*1998-99*	*1999-2000*	*2000-01*
Cotton textiles	125.6	115.9	123.7	127.3
Wool, Silk	172	176.8	197.8	209.3
Jute	114.3	106	105	105.8
Textile	158.7	153.1	156.1	162.4
Wood	128.5	121.0	101.4	104.3
Paper	146.4	169.8	180.5	164.0

Source: Economic Survey, 2000-01.

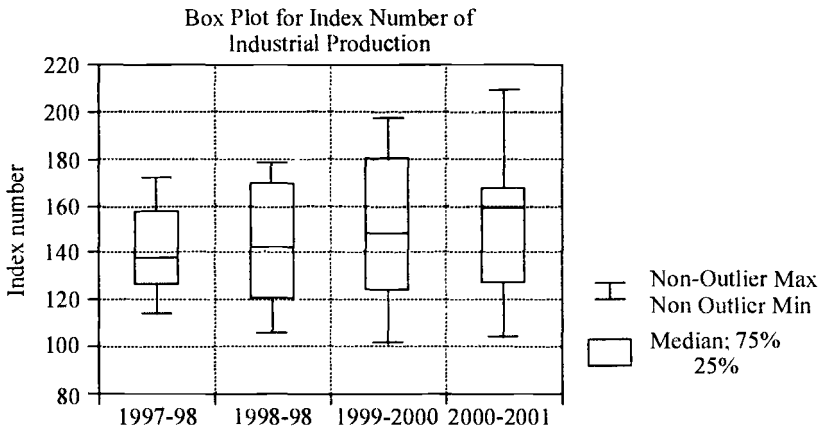


Fig. 14.1. *Box Plot for data provided in Table 14.1*

From the plot we can say that the indices of 2000-01 has the maximum scattering, the same year also shows considerable shift in the median compared to the other years. However, both the location and dispersion appears to be minimum in case of indices corresponding to 1997-98. Also, the indices are more asymmetric about the median in case of 1997-98 and 2000-01 compared to the other two years.

14.3. Axes

X-Axis: It can represent the values of the variable that we suspect may have a relation to the response variable. For this data set we consider years along the X axis.

Y-Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took the indices along the Y axis.

14.4. Advantages

(*a*) To compare the difference between the location (central value) of different series of observations.

(*b*) To compare the scattering of data in two or more sets of the observations.

(*c*) To understand how much symmetrical a particular data set is about its median.

(*d*) In the box plot the use of color is not very essential.

14.5. Disadvantages

(*a*) The box plot can be used as a handy tool for comparison, for different data sets only if the range of the data sets is not much apart and also if the data sets are measured in the same unit. In case they are far apart then it is better to draw a single box plot for each data set separately.

(*b*) The box plot is less attractive and needs a lot of computation before it is to be drawn.

14.6. Some Additional Insight

Slight variation in the box plot also provides options for the detection of outliers. Tukey (1977) used box plot for the identification of outliers. Filliben (1997) writes of a mean box plot, which is based on means and standard deviations instead of median and quartiles. McGill, Tukey and Larsen (1978) added the concept of notches in the box plot which can be used for the purpose of comparison of the medians in case of a multiple box plot. Also in the same paper the authors introduced the concept of use of the breadth of the box plot to represent the size of the groups in case of unequal group sizes.

14.7. Some Related Techniques

(*a*) Block Plot;

(*b*) Jittered Plot; and

(*c*) Analysis of Variance.

15. BOX-COX LINEARITY PLOT

15.1. Definition and Description

When performing a linear fit of Y against X, an appropriate transformation of X can often significantly improve the fit. The Box-Cox transformation introduced by Box and Cox in 1964 is a particularly useful family of transformations. The transformation is given by,

$$Z = \frac{X^{\lambda} - 1}{\lambda} \qquad ...(15.1)$$

where X is the variable being transformed, Z is the transformed variable and λ is the transformation parameter. If $\lambda = 0$, the natural log of the data is taken instead of using the above formula. The Box-Cox linearity plot is a plot of the correlation between Y and the transformed X *i.e.*, Z for given values of λ. That is, λ is the coordinate for the horizontal axis variable and the value of the correlation between Y and Z is the coordinate for the vertical axis of the plot. The value of λ corresponding to the maximum correlation (or minimum for negative correlation) on the plot is then the optimal choice for λ. Now the optimal value of λ thus obtained, can be used in (15.1) and accordingly the value of Z under the transformation may be obtained.

15.2. Working Data

The working data comprises of the Advances in crores[1] of rupees, provided by the various Regional Rural Banks of India for a time period of 1991–2004 in Table 15.1 below:

Table 15.1: Year-wise advances (in crores), provided by the various Regional Rural Banks of India

Year (X)	*Advances (Y)*
1991	3535.35
1992	4090.86
1993	4626.73
1994	5253.02
1995	6290.97
1996	7505.03
1997	8718.08
1998	9861
1999	11355.84
2000	12663

(contd...)

1. One Crore = 10 millions

Year (X)	*Advances (Y)*
2001	15579
2002	18373
2003	22157.85
2004	26112.86

Source: RBI Bulletin.

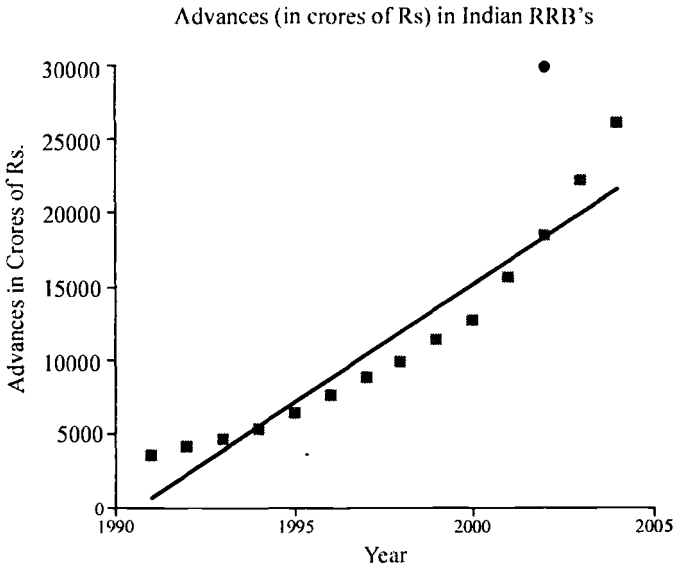


Fig. 15.1. *A linear fit to the original data*

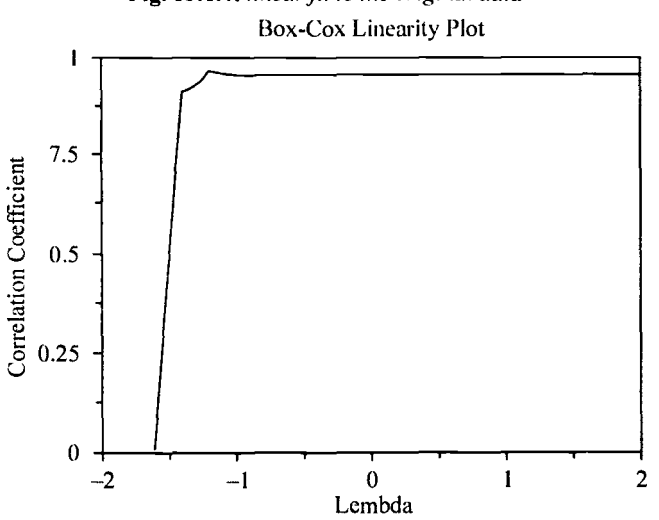


Fig. 15.2. *Box-Cox Linearity Plot to the Data in Table 15.1*

The plot shows that the maximum value of the correlation coefficient is attained for any value of $\lambda > -1$. *Thus, we may take* $\lambda = -1$.

Thus we have $$Z = \frac{X^{\lambda} - 1}{\lambda} = \frac{X^{-1} - 1}{-1} = 1 - \frac{1}{X} \quad ...(15.2)$$

as the transformation to which the values of X may be subjected and accordingly a much better liner fit will be obtained of the values of Z with that of Y (Advances).

15.3. Axes

X-Axis: It can represent the values of λ.

Y-Axis: The vertical axis consists of correlation between Y and the transformed X *i.e.*, Z for given values of λ.

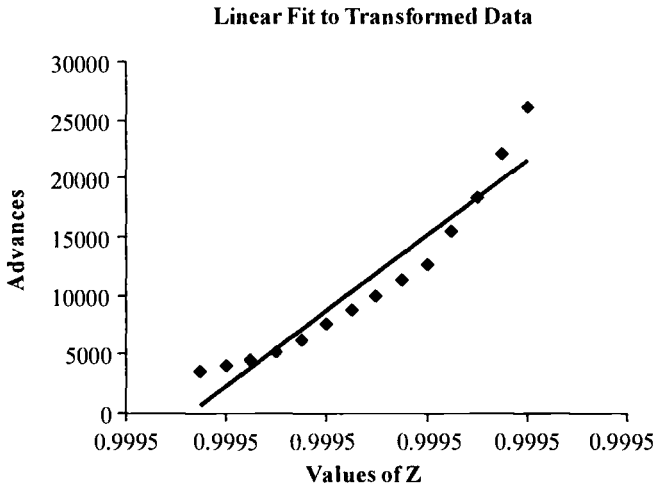


Fig. 15.3. *Plot of liner fit based on the transformed data*

15.4. Advantages

(*a*) The plot gives us an idea about the type of transformation that would improve the quality of fit.

(*b*) To determine the optimal value of the transforming parameter λ of the transformation equation.

15.5. Disadvantages

(*a*) The calculations involved are very lengthy.

(*b*) Not a very popular plot amongst the commonly used statistical packages.

(*c*) The transformation is not always very successful, as seen in this case.

15.6. Related Techniques

1. Linear Regression; and
2. Box-Cox Normality Plot.

16. BOX-COX NORMALITY PLOT

16.1. Definition and Description

The assumption of normality is the gate keeper of many statistical tests and estimations. But many real data sets are not approximately normally distributed. The Box-Cox transformation introduced by Box and Cox in 1964 can be used to transform the data into approximate normal form. The transformation is given by,

$$Z = \frac{X^{\lambda} - 1}{\lambda} \qquad \text{...(16.1)}$$

where X is the response variable being transformed, Z is the transformed parameter and λ is the transformation parameter. If $\lambda = 0$, the natural log of the data is taken instead of using the above formula.

Given a particular transformation, it is now necessary to measure the normality of the resulting transformation. One measure is to compute the correlation coefficient of a normal probability plot. The correlation is computed between the vertical and horizontal axis variables of the probability plot and is a convenient measure of the linearity of the probability plot. The Box-Cox normality plot is the plot of these correlation coefficients for different values of λ. The value of λ corresponding to the maximum correlation on the plot is then the optimal choice for λ. Now the optimal value of λ thus obtained, can be used in (16.1) and accordingly the normality to the non–normal data can be obtained.

16.2. Working Data

The sample data for the plot is taken from Anderson (1958). The data gives the measurement (in mm) of head length of 36 adult sons in a few families.

Table 16.1: Head length of sons of some families

191, 195, 181, 183, 176, 208, 189, 197, 188, 192, 179, 183, 174, 190, 188,
163, 195, 196, 179, 201, 185, 188, 171, 192, 190, 189, 197, 187, 186,
174, 185, 195, 187, 161, 173, 185.

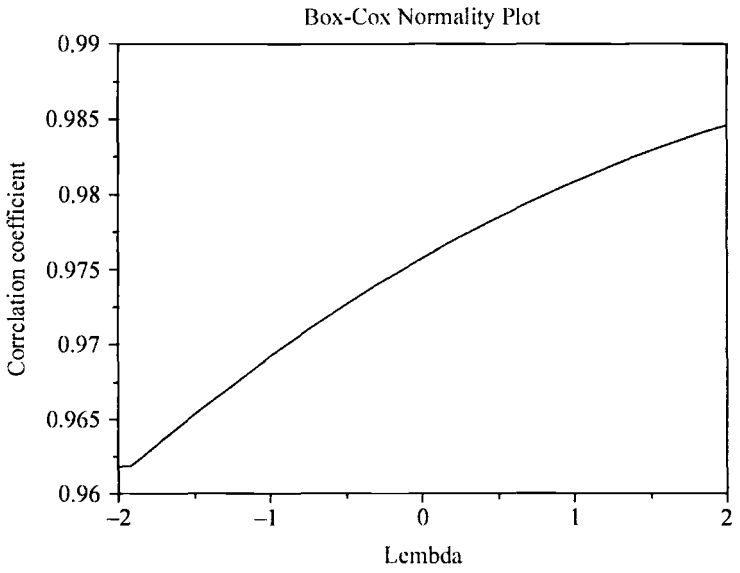


Fig. 16.1. *A Box-Cox Normality Plot*

From the plot we see that the maximum correlation is obtained for $\lambda=2$. *Thus the transformation obtained by putting* $\lambda=2$ *in (16.1) we will get an approximate transformation to normality of the data set.*

Thus the required transformation can be

$$Z=\frac{X^2-1}{2}$$

16.3. Axes

X Axis: It can represent the values of λ.

Y Axis: The vertical axis consists of correlation coefficient obtained from the normal probability plot after applying the said transformation for different values of λ.

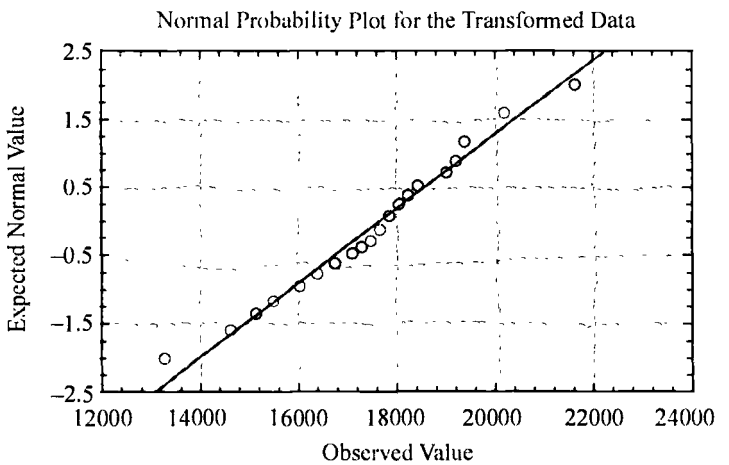


Fig. 16.2. *A normal-probability plot with the transformed data*

The plot shows that the transformation is partially successful in approximately converting the data into normal form.

16.4. Advantages

(*a*) The plot gives us an idea about the type of transformation that would normalize the data.

(*b*) To determine the optimal value of the transforming parameter *i.e.*, λ that would under go the transformation.

16.5. Disadvantages

(*a*) The calculations involved are very lengthy.

(*b*) Not a very popular plot amongst the commonly used statistical packages.

(*c*) The transformation is not always very successful.

16.6. Related Techniques

(*a*) Normal Probability Plot;

(*b*) Box-Cox Linearity Plot;

(*c*) Anderson-Darling Test;

(*d*) Wilks Shapiro Test; and

(*e*) EDF Plot.

17. BUBBLE PLOT

17.1. Definition and Description

The bubble plot is a graphical technique that can be used for representing three variables each of which takes numerical values. This plot can be considered as an extension of a scatter plot. Here we choose two primary variables that are represented along X-axis and Y-axis respectively. The (x, y) points are plotted in the graph paper but the plotting symbols are circles. The radii of the circles are proportional to the value of a third variable. Precisely speaking the bubble plots represent the values of three variables by drawing circles of varying sizes at points that are plotted on the vertical and horizontal axes. Two of the variables determine the location of the data points, while the values of the third variable control the radius of the circles. The circles thus formed are not filled so that they remain transparent and the overlapping of points can be easily understood. Once the plot is drawn, the plotted points appear like bubbles of varying size. This explains the reason why the plot is named so.

17.2. Working Data

The data used for drawing the bubble plots is given in Table 17.1. The data comprises of heights, weights, age and sex of 30 individuals collected non-randomly from some known individuals.

Table 17.1: Data of 30 individuals pertaining to their height, weight and age

Height (feet)	*Weight (kg)*	*Age (Years)*
5.17	50	22
5.67	54	24
5	45	21
5.5	60	26
5.08	38	13
5.25	55	51
5.58	51	48
5.41	53	15
5.75	62	27
5.66	72.5	21
5.17	42	15
5.58	58	61
5.08	41	54

Height (feet)	*Weight (kg)*	*Age (Years)*
5	40	21
5.67	69	21
4.67	39	24
5 25	57	28
6	70	26
5.58	65	28
5.42	50	21
5.17	47	55
5 33	60	52
5.5	64	58
5.83	56	21
4 75	50	12
5.25	72	40
5.5	73	31
5.58	71	61
5.08	54	59
5	57	28

The bubble plot in Figure 17.1 represents the first three variables of the Table viz. Height, Weight and Age.

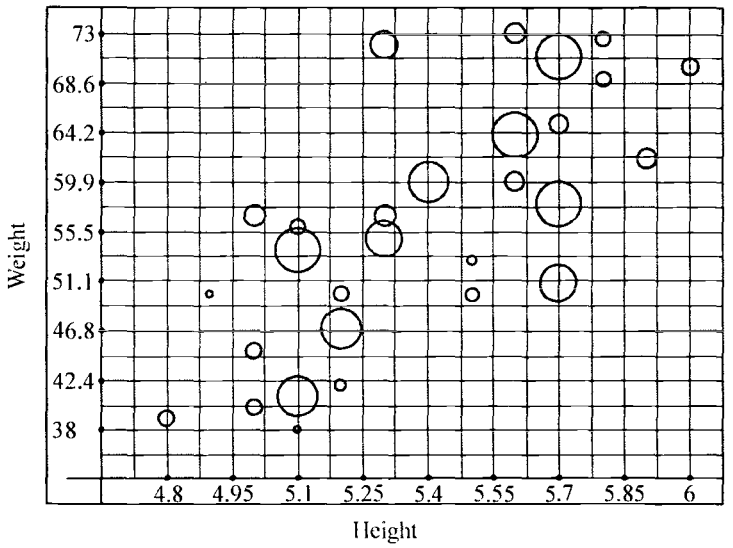


Fig. 17.1. *A bubble plot for the data in Table 17.1*

17.3. Axes

X Axis: It can represent the values of the variable that we suspect may have a relation to the response variable. For this data set we consider weight to be related to the height of the individual and hence we have taken height of the individual along the X-axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took weights along the Y-axis.

The third variable is age and is represented by the radius of the circle. Thus the radius of each of the plotted circle is kept proportional to the age of the subject.

17.4. Advantages

(*a*) The graphs can be used to study the relationship between X and Y variables.

(*b*) It can also be used to study the variation between X and Y in presence of a third variable.

(*c*) It is helpful in detection of outliers.

(*d*) It can be used for visualization of three variables in two dimensions.

17.5. Disadvantages

(*a*) This graphical technique requires some calculations for converting the value of the third variable to proportional radius.

(*b*) Bubble plot may sometime lead to over plotting. In case of many observations the bubbles may over lap either partially or entirely and hence make the representation untidy.

(*c*) The numerical variable that is represented by the radius of the plotted circle is studied relative to each other. The values of such variables are not scaled to be read from the graph and thus the graph remains less informative about the third variable.

17.6. Related Techniques

(*a*) Glyph Plot;

(*b*) Sunflower Plot; and

(*c*) Three Dimensional Scatter Plot.

18. BUBBLE PLOT (CATEGORICAL)

18.1. Definition and Description

This is an extension of the bubble plot. This plot is a graphical technique that can be used for representing four variables three of which takes numerical values and the fourth one is a categorical variable. The bubble plot is drawn first in the manner discussed in the earlier plot and then the boundaries of the bubble are colored differently based on the value of the categorical variable. Here different colors are used to represent the different categories under the fourth variable, and accordingly we get bubble of different colors and varying sizes in the plot. Thus this plot can be used to represent four variables– three numerical variables and the fourth one is a categorical variable.

18.2. Working Data

The data used for drawing the bubble plots is given in Table 18.1. The data comprises of heights, weights, age and sex of 30 individuals collected non-randomly from some known individuals. The table is an extended form of Table 17.1

Table 18.1: Data of 30 individuals pertaining to their height, weight, age and sex

Height (feet)	*Weight (kg)*	*Age (Years)*	*Sex*
5.17	50	22	F
5.67	54	24	M
5	45	21	F
5.5	60	26	M
5.08	38	13	M
5.25	55	51	F
5.58	51	48	M
5.41	53	15	F
5.75	62	27	M
5.66	72.5	21	M
5.17	42	15	F
5.58	58	61	M
5.08	41	54	F
5	40	21	F
5.67	69	21	M
4.67	39	24	F
5.25	57	28	M

(contd.)

Height (feet)	*Weight (kg)*	*Age (Years)*	*Sex*
6	70	26	M
5.58	65	28	M
5.42	50	21	F
5.17	47	55	F
5.33	60	52	F
5.5	64	58	M
5.83	56	21	M
4.75	50	12	M
5.25	72	40	F
5.5	73	31	M
5.58	71	61	M
5.08	54	59	F
5	57	28	F

The bubble plot in Figure 18.1 represents the four variables of the Table viz. Height, Weight, Age and Sex.

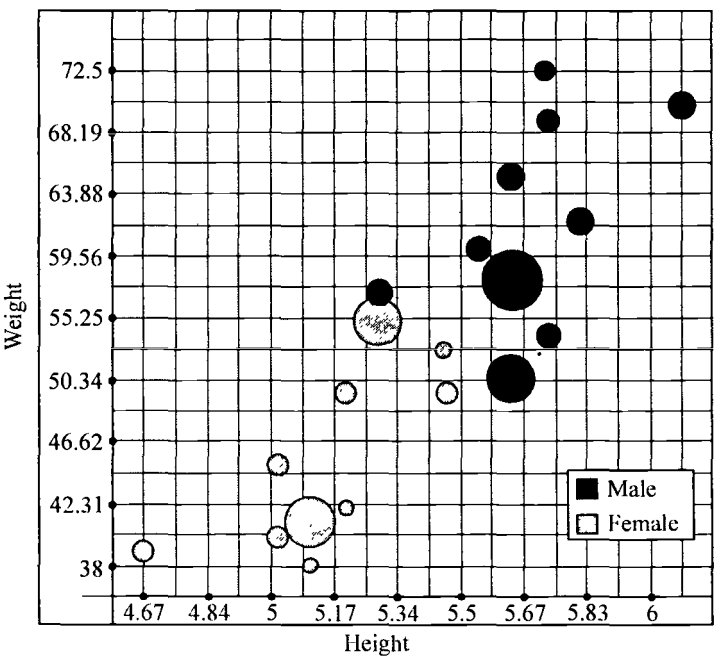


Fig. 18.1. *A categorical bubble plot for the data in Table 18.1*

18.3. Axes

X Axis: It can represent the values of the variable that we suspect may have a relation to the response variable. For this data set we consider weight to be related to the height of the individual and hence we have taken height of the individual along the X–axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took weights along the Y axis.

The third variable is age and is represented by the radius of the circle. Thus the radius of each of the plotted circle is kept proportional to the age of the subject.

The fourth variable is a categorical variable and is represented by the color of the bubble. Here a black bubble is used to represent the Males and the grey bubble is used to represent Females.

18.4. Advantages

(*a*) The graphs can be used to study the relationship between X and Y variables.

(*b*) It can also be used to study the variation between X and Y in presence of two other variables.

(*c*) It is helpful in the detection of outliers.

(*d*) It can be used for visualisation of four variables in two dimensions.

18.5. Disadvantages

(*a*) This graphical technique requires some calculations for converting the value of the third variable to proportional radius.

(*b*) Bubble plot may sometime lead to over plotting. In case of many observations the bubbles may overlap either partially or entirely and hence make the representation untidy.

(*c*) The numerical variable that is represented by the radius of the plotted circle is studied relative to each other. The values of such variables are not scaled to be read from the graph and thus the graph remains less informative about the third variable, 'Age' in this case.

18.6. Related Techniques

(*a*) Categorical Glyph Plot;

(*b*) Categorical Sunflower Plot; and

(*c*) Categorical Scatter Plot.

19. CARTOGRAM

19.1. Definition and Description

When data are to be displayed in geographical basis one takes the help of cartograms. This type of diagram is also known as statistical map. Cartograms are mainly used for comparing geographical data. For drawing this diagram, the map of the geographical area is first drawn and then the figures are represented by symbols of various shades, colors or sizes (where size of the symbols is proportional to numerical figures). Sometimes pictograms are also used to represent the quantities. In such cases different quantities are coded using different symbols and accordingly are plotted in the graph.

19.2. Working Data

The data used for drawing the cartogram is provide in the Table 19.1 which pertains to the production of Tea and Coffee in the various states of India in Lakh Tonnes in 1981–82.

Table 19.1: Production of Tea and Coffee in India during 1981-82

(in Lakh Tonnes)

State	*Tea*	*Coffee*
Assam	3.21	—
West Bengal	1.48	—
Tamil Nadu	0.67	0.14
Kerela	0.44	0.24
Karnataka	0.03	0.82

19.3. Axes

In this plot no axes are used.

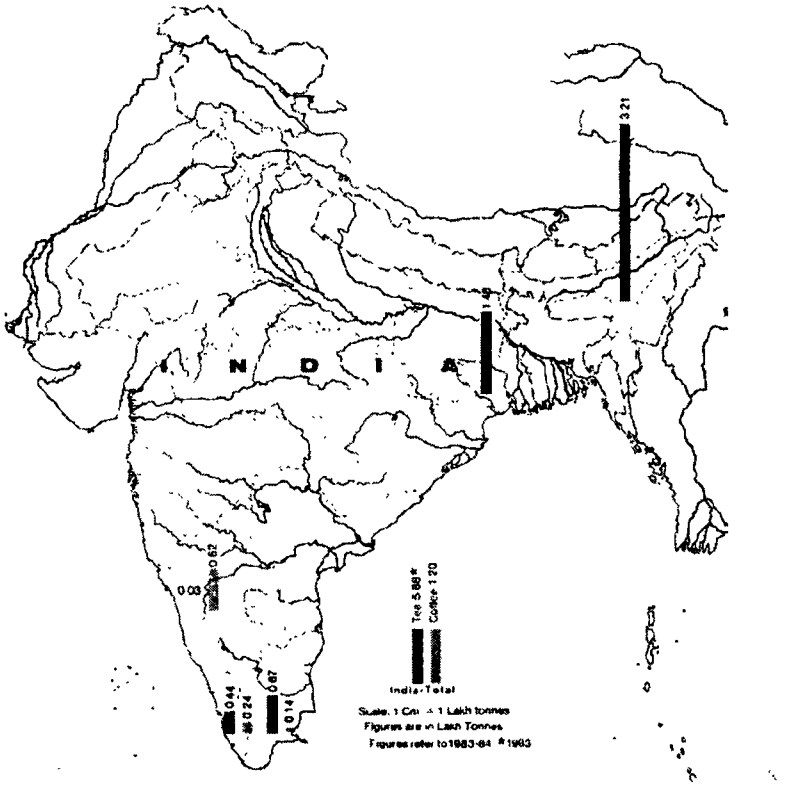


Fig. 19.1. *A cartogram to the data in Table 19.1*

This cartogram is scanned from, "The New Taj Mahal Atlas", Published by Orient Longman Ltd. Madras, India.

19.4. Advantages

(*a*) This plot does an excellent work of comparing the feature/ features geographically.

(*b*) Multiple number of variables can be represented for the various geographical areas.

(*c*) The plot is very popular and is used in various disciplines.

19.5. Disadvantages

(*a*) Special purpose software is required for producing such graphs. Commonly used statistical software does not provide this graph.

(*b*) Since the various statistical graphs superimposed in the map (bar in this case) are not drawn relative to any scale so these figures may be unreliable in cases.

19.6. Related Techniques

(*a*) Pictogram.

20. CATEGORICAL SCATTER PLOT

20.1. Description of the Plot

The categorical scatter plot is a graphical technique that can be used for representing three variables. Two of which are numerical values and the third one represents the categorical value. Here the two numerical values are represented along X-axis and Y-axis respectively. The categorical values are given some codes. This codes may be either numericals like 1, 2 etc. or alphabets like A, B etc. It is the code which is plotted in the graph corresponding to the point (x_i, y_i). Thus the graph looks like a scatter diagram with the points being replaced by corresponding codes of the categorical variable.

20.2. Working data

The data used for drawing the categorical scatter plot given in Table 20.1. The data comprises of heights, weights, and sex of 30 individuals.

Table 20.1: Data corresponding to height, weight and sex of 30 individuals

Height (feet)	*Weight (Kg)*	*Sex*
5.17	50	F
5.67	54	M
5	45	F
5.5	60	M
5.08	38	M
5.25	55	F
5.58	51	M
5.41	53	F
5.75	62	M
5.66	72.5	M
5.17	42	F
5.58	58	M
5.08	41	F
5	40	F
5.67	69	M
4.67	39	F
5.25	57	M
6	70	M
5.58	65	M

(contd...)

Height (feet)	*Weight (Kg)*	*Sex*
5.42	50	F
5.17	47	F
5.33	60	F
5.5	64	M
5.83	56	M
4.75	50	M
5.25	72	F
5.5	73	M
5.58	71	M
5.08	54	F
5	57	F

Note: While plotting female is represented as '1' and male as '2'.

20.3. Axes

X Axis: It can represent the values of the variable that we suspect may have a relation with the response variable. So here we take height along the X-axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. So we take weight along the *Y*-axis.

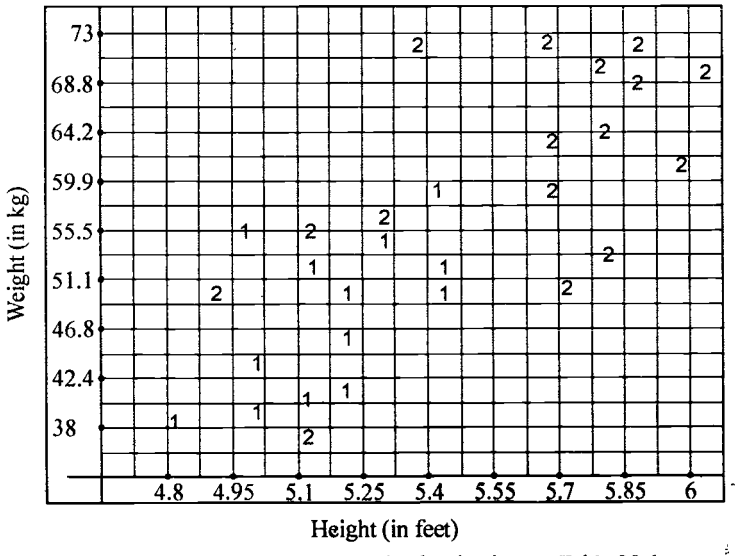


Fig. 20.1. *Categorical scatter plot for the data in Table 20.1*

From the plot it is clear that there is wide variation in the height of the males and females as the '1' and '2' forms separate clusters with a different horizontal as well as vertical alignment. This implies that the two sexes show difference in their average weight. Likewise, we may view the difference in the horizontal alignments of the two separate clusters for males and females and conclude that the two groups differ in their corresponding heights.

20.4. Advantages

(*a*) The graph can be used to study the relationship between X and Y variables.

(*b*) It can also be used to study the variation between X and Y in presence of a third variable. This plot enables us to visualize if there is any significant difference between the response variable or the independent variable across the various categories.

(*c*) It is helpful in detection of outliers.

(*d*) It can be used for visualization of 3 variables (two numerical variables and one categorical variable) in two dimensions.

20.5. Disadvantage

(*a*) In this plot three variables of which two are numerical variables and the third one has to be a categorical one.

(*b*) With the increase in the number of categories under the categorical variable the plot becomes clumsy as different plotting symbols will be used making it difficult to interpret.

20.6. Related Techniques

(*a*) Categorical Glyph Plot;

(*b*) Categorical Sunflower Plot; and

(*c*) Categorical Bubble Plot.

21. c-CHART

21.1. Definition and Description

c-Chart is a type of control chart used for attributes. It is the control chart for number of defects per unit. Control charts are used to study the variability in the quality of a manufactured product and are used to understand if the process is within control. If the quality is not directly measurable but the number of defects in a particular unit can be measured then this type of control chart can be used.

Here k samples of the manufactured product are collected of same size 'n' (say). Let c_i be the number of defects in the i^{th} ($i = 1, 2, \ldots, k$) sample. For a quality product the number of defects in an observation is supposed to follow Poisson distribution. If the population value of the parameter (λ) is not known then it can be estimated by $\bar{C}$, where

$$\bar{C} = \frac{1}{k}\sum c_i$$

A control chart consists of three lines parallel to the X axis, the control line (CL), the upper control limit (UCL) and the lower control limit (LCL). Here, we have

Control Line $= CL = \bar{C}$

Upper Control Limit $= UCL = \bar{C} + 3\sqrt{\bar{C}}$

Lower Control Limit $= LCL = \bar{C} - 3\sqrt{\bar{C}}$

After all this calculations are done the values of c_i are plotted for different values of i. In other words the values of the number of defects in the different samples are plotted in the form of dots against the sample numbers. The UCL and LCL are also drawn. If all the dots *i.e.*, (i, c_i) fall within the UCL and LCL (control limits) then the system is said to be under control otherwise the system is beyond control.

21.2. Working Data

The data for this purpose is generated from 16 boxes containing electric switches where each box contains 20 switches. The boxes are selected randomly from a large consignment and are inspected for the number of defects per box. Accordingly the following data is obtained.

Table 21.1: Number of defects in the different boxes full of switches

Box No.	*No. of Defects*	*Box No.*	*No. of Defects*
1	12	9	11
2	15	10	12
3	9	11	16
4	14	12	13

(contd...)

Box No	No. of Defects	Box No.	No. of Defects
5	18	13	19
6	26	14	18
7	8	15	14
8	6	16	21

Here, $\bar{C} = \frac{\sum c_i}{k} = \frac{232}{16} = 14.5$

Upper Control Limit = $UCL = \bar{C} + 3\sqrt{\bar{C}} = 14.5 + 3\sqrt{14.5} = 25.93$

Lower Control Limit = $LCL = \bar{C} + 3\sqrt{\bar{C}} = 14.5 - 3\sqrt{14.5} = 3.07$

Control Limit = $CL = \bar{C} = 14.5$

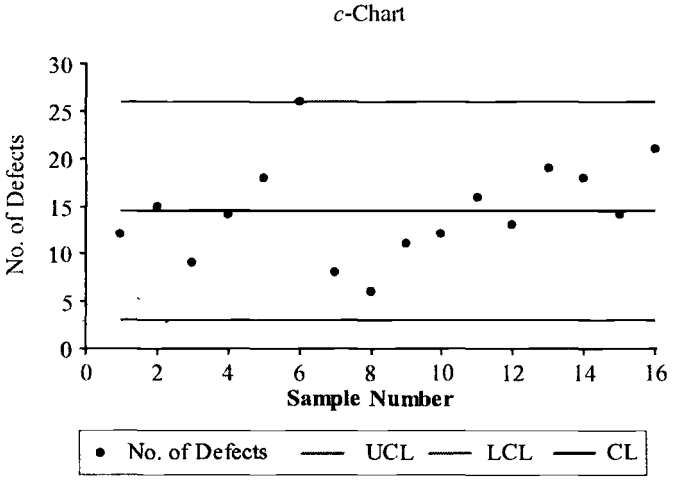


Fig. 21.1. *c chart for the data in Table 21.1*

The control chart shows that the number of defects in one of the samples is outside the control limits and accordingly we may comment that the system is out of control.

21.3. Axes

X Axis: The axis is used to represent the sample numbers.

Y Axis: This axis is used to represent the values of the number of defects in the sample.

21.4. Uses

(*a*) The plot is simple to draw and the calculations are relatively simple.

(*b*) The chart is used to discover the existence of any assignable cause of variation.

(*c*) The chart is successful even in case of small samples.

(*d*) The chart can deal with situations where the quality of the product cannot be measured but is only an attribute.

21.5. Related Techniques

(*a*) Process Control;
(*b*) Control Chart;
(c) *s* Chart;
(*d*) $\overline{X}$ Chart; and
(*e*) *p* Chart.

22. CHERNOFF'S ICON PLOT

22.1. Definition and Description

Chernoff (1973) proposes a very interesting icon for representing multivariate data. He associates each variable with different characteristics of human face for example eyes, ears, lips, mouth etc. With change in the value of the variable, the attribute with which it is attached also changes. This is a very complex type of icon and a relatively complicated computer program should be developed in order to draw the plot. However, some of the statistical software supports the plot, and on using the software the plot is relatively easier to generate. The initial plot was drawn by Chernoff consisted of 8 variables each being represented by a particular feature of the face like length of the nose, size of the eyes, curvature of the mouth etc. Thus with change in the value of any one of the variable a change takes place in the corresponding feature of the face and so each icon can be differentiated easily. A comment of Kotz and Johnson (1982) seems to be very relevant– "The cartoons of faces seem to be very effective for this purpose. People grow up studying and reacting to faces. Small and barely measurable differences are easily detected."

Several authors has brought about various modifications to the original faces drawn by Chernoff amongst which the most remarkable change was brought about by Marshall (1974) who used the cartoon figure of a footballer, where the shape of the various parts of the body of the footballer was used to present the strength and weakness of a football team.

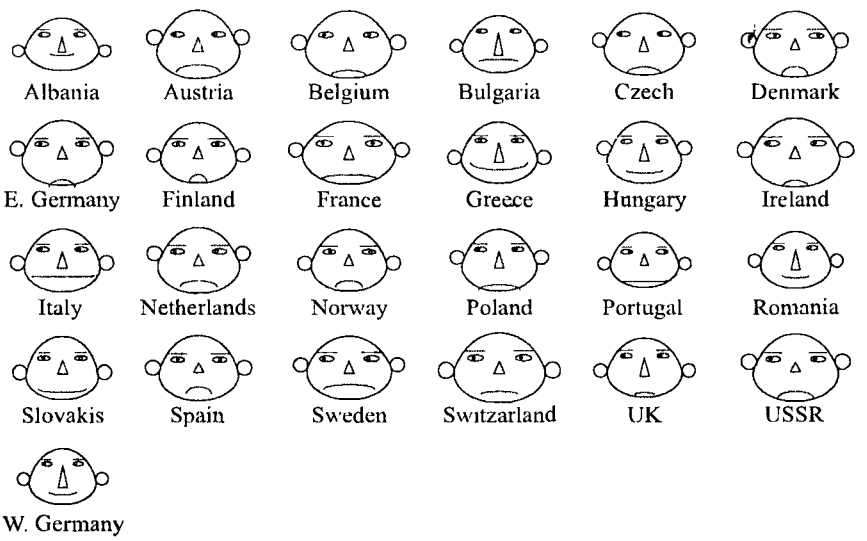
22.2. Working Data

The data used for this purpose is provided in Table 22.1 below. The data set is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups is given.

Table 22.1: Protein consumption in European countries

	Red meat	*White meat*	*Eggs*	*Milk*	*Fish*	*Cereals*	*Starchy Food*	*Pulses*	*Fruits Veg.*
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19.9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
East Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9 5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4

	Red meat	*White meat*	*Eggs*	*Milk*	*Fish*	*Cereals*	*Starchy Food*	*Pulses*	*Fruits Veg.*
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	45.7	2.2	7.8	6.5
Hungary	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherland	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4 7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1 1	4 9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Slovakia	6.2	7.2	1.5	13.4	3	29	4.2	5.3	2.7
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzarland	13.1	10.1	3.1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
West Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1 5	3.8



Lengend: face/w = R_MEAT, ear/lev = W_MEAT, halfface/h = EGGS, nose/l = CEREALS mouth/cent = ST_FOOD, mouth/curv = PULSES, mouth/l = FRUITS, pupils/pos = FISH, eyebrows/l = MILK

Fig. 22.1. *The Chernoff's icon plot for the protein consumption data*

22.3. Axes

In this plot no axes are used.

22.4. Advantages

(*a*) The faces have become very popular for representation of multivariate data and have been used for variety of applications including the study of moon rocks, medical science, business data etc. (Kotz and Johnson: 1982).

(*b*) Experiences with other icons tell that an important aspect of icon displays is partitioning of the observations into different clusters by visual detection only. This is easier to perform with Chernoff faces as, similar type of faces can be easily identified and this can be done even by a novice.

(*c*) In case of Chernoff's icon plot an additional variable will lead to an additional feature of the face to come into act in the plot. This will not occupy any extra space unlike other icon plots like column icon plot or profile plot etc.

(*d*) Software like, STATISTICA, Systat have the option of drawing such a plot.

22.5. Disadvantages

(*a*) Cleveland and McGill (1984) expressed their views against the use of Chernoff faces as they think that for an icon those symbols should be used that facilitate mental conditioning and have relatively high perceptual accuracy of extraction. The preference to icon plots must thus go to those symbols which have stereo depth, orientation, linear size and color. For example, if the position of the ears in a Chernoff face represents a variable, it is difficult to focus attention on all the faces and detect the trend or relation with other variables.

(*b*) It is impossible to draw the plot manually as the small changes of facial expressions and relating it to the various numerical values are impossible to bring about. Only related software or a high level graphical program can draw the plot.

(*c*) The visual detection of the values of the variable is a very difficult task from faces even if the receiver is trained to interpret the data from a graph.

22.6. Related Techniques

(*a*) Profile Icon Plot;

(*b*) Star Icon Plot;

(*c*) Chernoff Faces;

(*d*) Column Icon Plot.

23. CHI PLOT

23.1. Definition and Description

Fisher and Switzer (1985) put forwarded a plot termed as the chi–plot that can be used to detect the presence of dependence between two variables and also the pattern of dependence in the data if any. Let (x_1, y_1), (x_2, y_2), ...,(x_n, y_n) be a random sample from H, the joint (continuous) distribution function for a pair of random variables (X, Y) and let $I(A)$ be the indicator function of the event A. For each data point (x_i, y_i), set

$$H_i = \sum_{j \neq i} I(x_j \leq x_i, y_j \leq y_i)/(n-1) \quad ...(23.1)$$

$$F_i = \sum_{j \neq i} I(x_j \leq x_i)/(n-1) \quad ...(23.2)$$

$$G_i = \sum_{j \neq i} I(y_j \leq y_i)/(n-1) \quad ...(23.3)$$

and

$$S_i = \text{sign}\left\{\left(F_i - \frac{1}{2}\right)\left(G_i - \frac{1}{2}\right)\right\} \quad ...(23.4)$$

On finding the values of H_i, F_i, G_i we combine them to find the values of χ_i using the following transformation,

$$\chi_i = \frac{H_i - F_i G_i}{F_i(1-F_i)\,G_i(1-G_i)} \quad ...(23.5)$$

and

$$\lambda_i = 4S_i \max\left\{\left(F_i - \frac{1}{2}\right)^2, \left(G_i - \frac{1}{2}\right)^2\right\} \quad ...(23.6)$$

Then the Chi-plot is drawn which is basically a scatterplot between (λ_i, χ_i), provided $|\lambda_i| < 4\left(\frac{1}{n-1} - \frac{1}{2}\right)^2$. The value of λ_i is a measure of the distance of the data point (x_i, y_i) from the center of the data set $(\tilde{x}, \tilde{y})$, where $\tilde{x}$ and $\tilde{y}$ are the medians values of (x_i, y_i), $i = 1, 2, \ldots, n$ respectively.

It is always helpful if the chi–plot is accompanied by a pair of horizontal grid lines at $\chi = -c_p/\sqrt{n}$ and $\chi = c_p/\sqrt{n}$, here c_p is selected such that approximately $100p\%$ of the pairs (λ_i, χ_i) lie between the lines. The two lines acts as confidence bands to the chi–plot and can be used as a measure of independence. Fisher and Switzer (1985) found the c_p values for $p = 0.9, 0.95$ and 0.99 and the values are, 1.54, 1.78 and 2.18 respectively. Other values of c_p for different levels of significance can be derived using the Monte Carlo techniques. From the discussion it can be understood that the chi–plot depends only on the data through their ranks, where the plotted points are such that they should be approximately horizontal under independence.

23.2. Working Data

The data used for drawing this plot is a hypothetical one showing the values of two variates say X and Y. There are 28 pairs of observations. They are provided in the table below.

Table 23.1: Some hypothetical values of X and Y

X	32	56	67	3	43	21	43.9	89.12	75	57	34	22	54	52
Y	12	32	45	5	7	67	78	86	54	55	43	32	43	21
X	31	24	44	41	62	32	37	58	89	43	56	76	79	88
Y	43	56	79	21	43	23	78	98	75	65	54	32	33	33

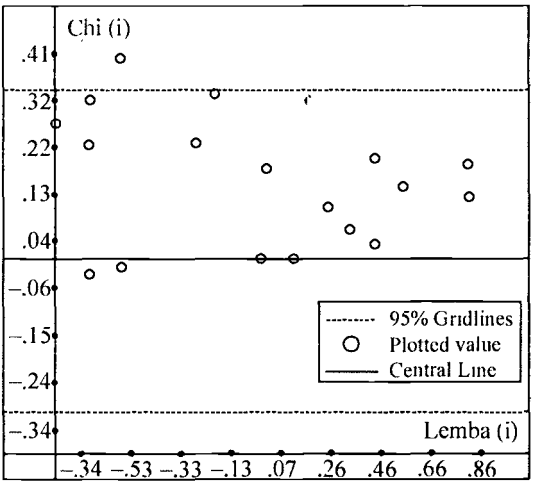


Fig. 23.1. *A chi plot for the data in Table 23.1*

From the plot it can be inferred that the variables are dependent on each other as some points of the chi plot falls beyond the 95% grid lines. The decision can be taken with 95% confidence.

23.3. Axes

X Axis: It can represent the values of the $\lambda_i = 4S_i \max\left\{\left(F_i - \frac{1}{2}\right)^2, \left(G_i - \frac{1}{2}\right)^2\right\}$, where F_i, G_i etc. are obtained using the equations (23.1), (23.2) and (23.3) etc.

Y Axis: The vertical axis consists of the values of $\chi_i = \frac{H_i - F_i G_i}{F_i(1-F_i)G_i(1-G_i)}$, where F_i, G_i etc., are obtained using the equations (23.1), (23.2) and (23.3) etc.

23.4. Advantages

(*a*) Provides a visual test for the dependence between two variables.

(*b*) Provides confidence intervals for such dependence which enables the viewer to differentiate if the dependence is significant for a given level.

(*c*) Can be used to study the randomness in a data set.

(*d*) Can be used to check if the residuals from a fitted model is randomly distributed and accordingly reach conclusions related to goodness of fit.

23.5. Disadvantages

(*a*) The calculations required for the plot is much complicated.

(*b*) Commonly used statistical software like SPSS, Systat, Statistica, etc. does not provide the option of drawing this plot.

(*c*) The test can be applied to large samples only.

23.6. Related Techniques

(*a*) Run Test;

(*b*) Serial Correlation Test; and

(*c*) Run Sequence Plot.

24. CHIGRAM

24.1. Definition and Description

Chigram is a graphical display technique used to differentiate between observed and fitted values through histograms. When histogram of a frequency distribution is drawn we can have a rough idea about the form of the probability distribution. If the plot is to check the normality of the data, and even if the histogram shows a bell shaped distribution, one cannot be sure of the normality of the data. This is because the normal curve is not the only bell shaped curve. To evaluate the distribution exactly, a comparison must come up between the observed frequencies and expected frequencies fit by the normal distribution. Let n_i be the observed frequency and $\hat{n}_i$ be the corresponding expected frequency. Now $n_i - \hat{n}_i$ may considered as the difference between observed and expected frequencies. But since the class frequencies are less in the tails so large differences between the observed and expected are noticed in the tails in case $n_i - \hat{n}_i$ are used. A chigram is a plot where $(n_i - \hat{n}_i)/\sqrt{\hat{n}_i}$ is plotted against each class. This has an advantage over the plotting of $n_i - \hat{n}_i$ is that it automatically acts as the variance stabilizer.

24.2. Working Data

For plotting the above-mentioned plots we take the data set from Gupta (1952) which is provided in Table 24.1. The data set showing the lifetime (in hours) of 300 electric lamps which follows normal distribution.

Table 24.1: Lifetime (in hours) of 300 electric lamps from Gupta (1952)

Life time (hours)	*Frequency*
950–1000	2
1000–1050	2
1050–1100	3
1100–1150	6
1150–1200	7
1200–1250	12
1250–1300	16
1300–1350	20
1350–1400	24
1400–1450	27
1450–1500	29
1500–1550	29
1550–1600	28

(contd...)

Life time (hours)	*Frequency*
1600–1650	25
1650–1700	21
1700–1750	16
1750–1800	12
1800–1850	8
1850–1900	6
1900–1950	3
1950–2000	2
2000–2050	1
2050–2100	1

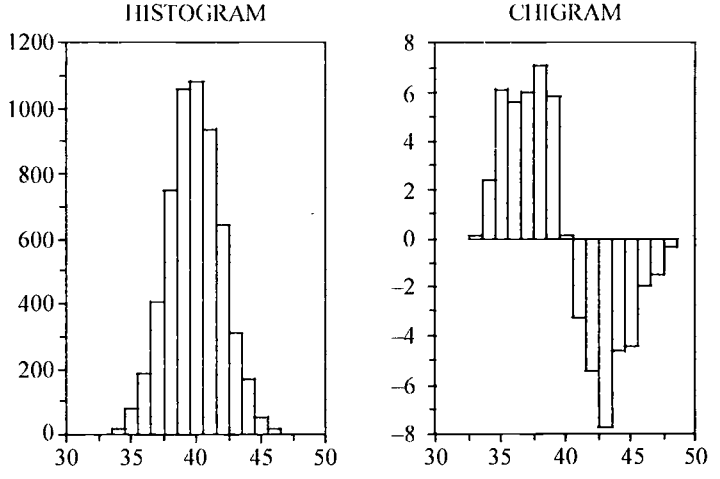


Fig. 24.1. *Histogram and chigram based on data in Table 24.1*

From the histogram it appears that the data follows a normal distribution but the chigram reveals that even the standardized residuals follow more or less uniform distribution in the first half and follows approximately normal distribution in the second half.

24.3. Axes

X Axis: It is used to represent the class intervals. For this data set Life Time (in hrs.) is taken along the X–axis.

Y Axis: The vertical axis is used for representing the standardized residuals *i.e.*, the values of $\frac{n_i - \hat{n}_i}{\sqrt{\hat{n}_i}}$ are taken.

24.4. Advantages

(*a*) They are very simple technique to judge the goodness of fit of a data set.

(*b*) The calculations involved are relatively simple. The can be easily understood and easily interpreted.

24.5. Disadvantages

(*a*) It is difficult to infer in all cases about the goodness of fit.

(*b*) The plot cannot be drawn for open-end classes.

24.6. Related Techniques

(*a*) Histogram;

(*b*) Residual Rootogram; and

(*c*) Residual Histogram.

25. CIRCLES

25.1. Definition and Description

Circles are two dimensional diagrams. In this plot circles are drawn to represent numerical values such that the area of the circles are proportional to the values it represent. We know that the area of the circle is proportional to the square of the radius of the circle. Thus in order to draw this plot initially the square root of the response variables is calculated. Then circles are then drawn with radius proportional to the value of these square roots.

25.2. Working Data

The data used for the plot is a hypothetical figure representing sale of a particular product launched by a company corresponding to the first four months are given below.

Table 25.1: Sale of a particular product in '000.

Month	*Sale (in '000)*	*Square Roots*
1	12	3.46
2	70	8.37
3	150	12.24
4	270	16.43

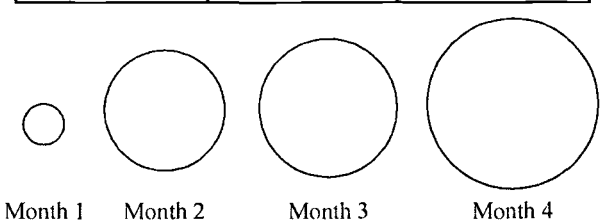


Fig. 25.1. *Circles corresponding to data in Table 25.1*

25.3. Axes

The diagram does not require the help of any axes.

25.4. Uses

(*a*) The plot is used for the representation of values of one variable only.

(*b*) The plot provides the option for two dimensional viewing of a set of data.

25.5. Related Techniques

(*a*) Sphere;

(*b*) Pyramid;

(*c*) Cone; and

(*d*) Cylinder.

26. COLUMN ICON PLOT

26.1. Definition and Description

Column or Histogram icon plots are icon plots used to display multivariate observations. It represents each observation by a set of p vertical bars. The height of any of these bars is proportional to the value of the variable it represents. Thus for plotting one observation having p variates, p bars are used. Freni-Titulauer and Louv (1984) pioneered this plot. In some cases the bars of the histogram are not attached to each other but are slightly separated. Such a plot is called as column icon plot. Each column of the plot is generally differently colored such that the variables can be easily identified. The same variable uses the same color for all the observations.

The bars across the observations can be compared for a particular variable, but the bars of the particular observation should not be compared to each other as they have different scale or different units of measurement. Henry (1995) suggests that one may standardized all the variables and then draw the plot. This would enable the user to rescale all the variables in a common scale and hence compare the performance of all the variables within an observation as well.

26.2. Working Data

The data used for this purpose is provided in Table 22.1. The data set is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups is given.

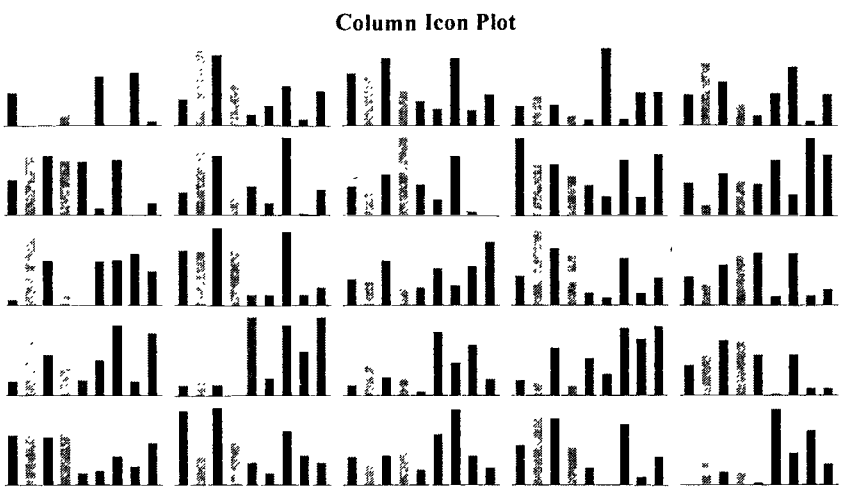


Fig. 27.1. *A column icon plot for the protein consumption data*

The icons are arranged in the order in which the data is presented. Since the data pertains to several European countries so one may arrange it region–wise. For example, East European countries, central European countries and so on. This would enable an onlooker to understand if the pattern of protein consumption varies in the different regions of Europe.

26.3. Axes

In this plot no axes are used. It is the order of the variables that are considered in a particular histogram. The variables are considered either from left to right or from right to left.

26.4. Advantages

(*a*) A histogram icon plot is easier to interpret.

(*b*) It is an excellent tool for visual detection as the heights of the bars can easily be compared.

(*c*) An important aspect of icon displays is partitioning of the observations into different clusters by visual detection only. This is not very easy but can be performed with histogram icon plot.

(*d*) The display does not require the use of color. STATISTICA, S-plus, Stat Graphics have the option of drawing such a plot.

26.5. Disadvantages

(*a*) Here it may be noted that a full set of bars in a histogram icon/column icon depicts an observation's performance across the multiple variables. Here each bar in an observation and is drawn with a different scale which depends on the range of the variables. The bars across the observations can be compared for a particular variable, but the bars corresponding to a particular observation should not be compared to each other as they have different scale.

(*b*) The column plot is less attractive in appearance compared to the other icon plots.

(*c*) It is difficult to determine the trends in the data from this display.

(*d*) With increase in the number of variables the space required for display of an individual observation increases in case of a histogram plot and hence it in not advantageous to represent more than five variables in those plots.

26.6. Related Techniques

(*a*) Star Icon Plot;

(*b*) Sunray Icon Plot ;

(*c*) Chernoff Faces;

(*d*) Pie Icon Plot; and

(*e*) Profile Icon Plot.

27. COLUMN PLOT (CIRCULAR BASE)

27.1. Definition and Description

This plot is similar to the multiple bar diagram but the base of the bars in no longer X axis but the circumference of a circle from which the bars come out in different directions, such that the length of the bars are proportional to the value it represents. Being placed in the circular base with the bars extending redially, the bars take a conic shape. Several concentric circles are drawn that acts as gridlines for the plot. These bars are equally spaced around the central point. The different bars representing different variables may be differently colored under each category. Such use of color makes it easier to identify a particular variable in a particular category.

The main purpose of the graph is to visualize the performance of different variables under different categories. The main problem with the usual multiple bar diagram is that with increase in the number of categories the bars keep on increasing and so more space is required. It often becomes difficult to space all the bars along the X axis. However this diagram having a circular base does not demand more area for the representation of any additional category.

27.2. Working Data

The data for this purpose and the plot is taken from the web site **www.mediaware-infotech.com.** The data relates to the rate of inflation of advertisement and circulation growth of various Hindi dailies of India.

Table 27.1: Rate of advertisement inflation and circulation growth in Hindi dailies

Newapaper	*Advertisement Inflation Rate*	*Circulation Growth*
Amar Ujala	6 83	26.68
Hindustan	20.6	9.67
Dainik Jagaran	13	NA
NavBharat Times	13.73	5.28
Dainik Bhaskar	3.92	5.83
Rajasthan Patrika	10.39	9
Aaj	5 63	–2.46
Dainik Jagaran (UP)	13.28	21.75

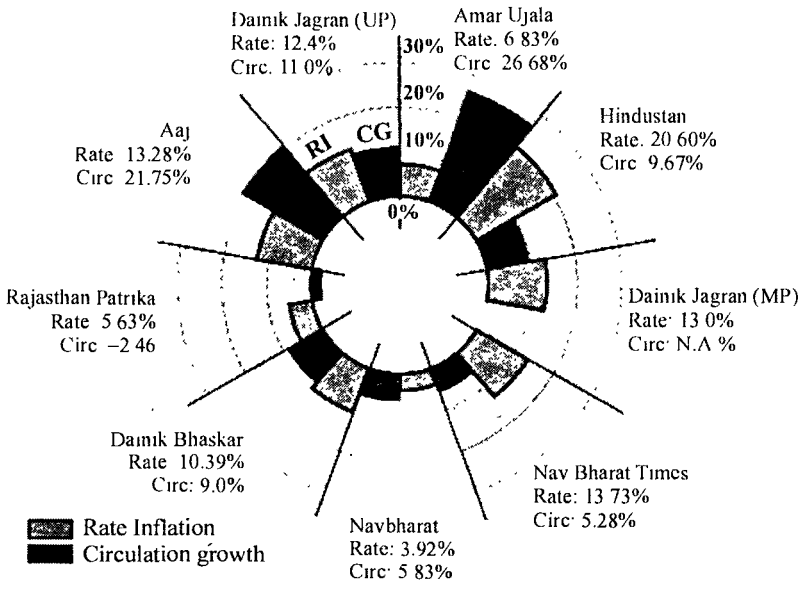


Fig. 27.1. *Column Plot of data provided in Table 27.1*

Notice the way in which negative values and missing values are plotted.

27.3. Axes

This plot does not require any axes for its drawing, only a radial scale is considered that can be demarked using concentric circles.

27.4. Advantages

(*a*) The plot is easy to interpret.

(*b*) The plot occupies less space compared to an equivalent multiple bar diagram.

27.5. Disadvantages

(*a*) The plot is difficult to draw manually.

(*b*) Most statistical software does not provide the option of drawing this plot.

(*c*) Large data sets are difficult to represent.

27.6. Related Techniques

(*a*) Column Chart;

(*b*) Multiple Bar Diagram; and

(*c*) Sunray Icon Plot.

28. COMOVEMENT PLOT

28.1. Definition and Description

The plot can be used for the purpose of plotting the comovement coefficients for a particular time series for different lags or for two time series for different lags. The plot consists of the comovement coefficient along the Y axis and the lag along X axis.

The comovement coefficient is similar to the formula of correlation coefficient only instead of subtracting the mean from x_i and y_i, it is x_{i-k} and y_{i-k} that is been subtracted respectively. The coefficient is given by,

$$C_1 = \frac{\sum_{i=2}^{n}(x_i - x_{i-1})(y_i - y_{i-1})}{\sqrt{\sum_{i=2}^{n}(x_i - x_{i-1})^2 \sum_{i=2}^{n}(y_i - y_{i-1})^2}} \text{ for lag 1.}$$

$$C_2 = \frac{\sum_{i=3}^{n}(x_i - x_{i-2})(y_i - y_{i-2})}{\sqrt{\sum_{i=3}^{n}(x_i - x_{i-2})^2 \sum_{i=3}^{n}(y_i - y_{i-2})^2}} \text{ for lag 2 and so on for lag } k.$$

The computed value of the coefficient lies between –1 and +1 and is useful in the study of comovement between two arbitrary time sequences.

28.2. Working Data

The table given below is taken from the Economic Survey, 2000-01 and shows the per capita availability of food grains net availability from 1990 to 2000.

Table 28.1: Net availability of cereals and pulses

Year	*Per capita availability per day (gms.)*	
	Cereals	*Pulses*
1990	435.3	41.1
1991	468.5	41.6
1992	434.5	34.3
1993	427.9	36.2
1994	434.5	37.2
1995	457.6	37.8
1996	443.4	32.8
1997	468.2	37.3
1998	417.3	33
1999	433.5	36.9
2000	426	32

Table 28.2: Comovement coefficients for different lags

Lag	1	2	3	4
Comovement Coefficients	0.7495	0.3214	0.6277	0.4632

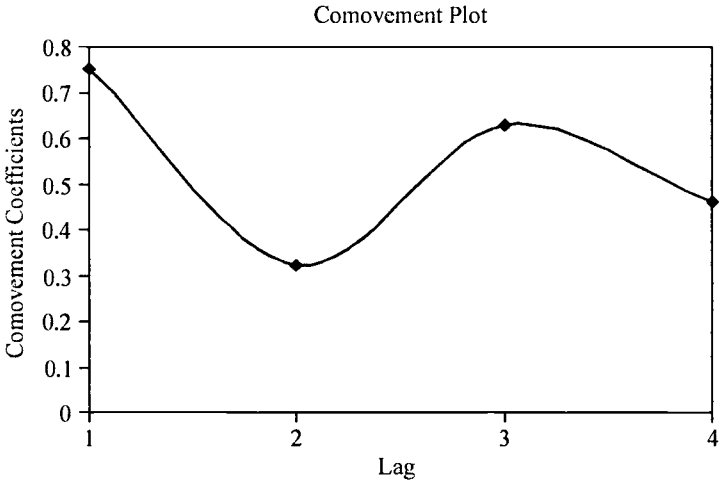


Fig. 28.1. *An Comovement Plot for the data in Table 28.2*

28.3. Axes

X Axis: In the axis the lags are considered. So the value starts from 1 and may go at least up to k, where $k = \frac{n}{4}$, converted to nearest integer in case k is fractional.

Y Axis: In the vertical axis we consider the comovement coefficients for different lags. So the maximum range of the axes is from -1 to $+1$.

28.4. Uses

(*a*) To visualize the comovement coefficients for between two time series for different lags.

(*b*) To provide an idea about the type of model that is appropriate for the data.

28.5. Some Related Techniques

1. Cross Correlation Plot;
2. Comovement Coefficient;
3. Serial Correlation Plot; and
4. Correlation Coefficient.

29. CONTOUR PLOT

29.1. Definition and Description

A contour plot is a graphical representation of the relationship between a response variable (Z) and two explanatory variables (X and Y). The Z values are plotted in the form of constant slices, called contours in the plane of the explanatory variables. The contour levels are plotted either as curves or as filled areas. A contour plot provides a "topographic map" of a function. In very simple words it is a projection of a three dimensional surface in a two dimensional plane. The contours join points on the surface that have the same height. It is better to have contours corresponding to an equally spaced z values. If one wants to use color in the contour plot then different colors may be used for different values of Z lying in the different ranges. Also if shades are used instead of color then one may use the shades in such a way that regions with higher z values are lighter.

29.2. Working Function

The function used for drawing the contour plot be:

$$Z = f(x, y) = \frac{x^2}{2} + y^2 \text{ with } x \in [-10, 10] \text{ and } y \in [-10, 10]$$

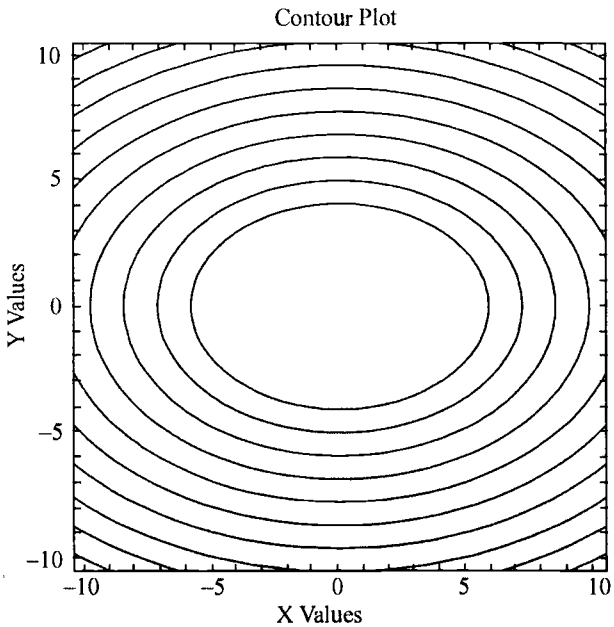


Fig. 29.1. *A contour plot for the above function*

From the figure we can make out that the minimum value of Z is at x = 0 and y = 0, and the value of the function is directly proportional to the absolute value of x or y irrespective of the sign of the value it takes.

29.3. Axes

X Axis: The first independent variable is considered along the axis.

Y Axis: The vertical axis is used for the representation of another independent variable.

29.4. Advantages

(*a*) To visualize a three dimensional plane in a two dimensional surface.

(*b*) For large data sets such plot gives us a guideline in understanding the relationship between variables.

(*c*) The plot can be used in designs of experiment to understand the difference between several treatments.

29.5. Disadvantages

(*a*) The iso-response values *i.e.*, the values of Z that can be connected to each other are very difficult to find out manually.

(*b*) Even if the iso-response values are obtained using the computer program the graph is difficult to draw manually.

(*c*) The graph is sometimes very difficult to interpret.

30. COPLOT

30.1. Definition and Description

Coplot is an abbreviation for conditioning plot. This technique is helpful in detecting the relationship between two variables when the effect due to other variables are kept constant. Some extensions of the conditional displays are the casement display and the trellis display. The coplots were initiated by Cleveland (1993). For a simple understanding of the coplot, let us consider three variables X, Y and Z (say). If Z is the conditioning variable then at the very outset this variable is divided into several subintervals in case Z is of continuous type. Now X and Y are plotted in the form of scatter plot, or any other suitable plot for various levels of Z. The plot is generated in the form of panels, where the first panel is used to show a series of intervals of the conditioning variable Z. This panel is called as the given panel. The horizontal axis of this panel corresponds to the conditioning sub–intervals of the conditioning variable, and the horizontal bars show the sub–intervals themselves. The sub–intervals are made in such a way that the number of observations within each sub–intervals is same, so the length of the sub–intervals will vary depending the data density. The panels below the given panel are called as the dependence panel. The dependence panel consists of scatter plots of the two other variables for various levels of Z. The tic mark and labels of the axes are plotted at alternate rows and columns.

30.2. A Sample Plot

The figure below is an example of a coplot scanned from Jacoby (1998).

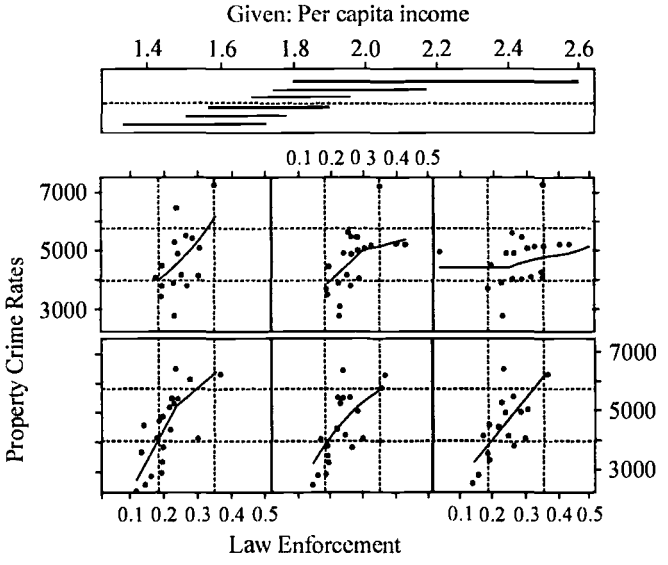


Fig. 30.1. *A coplot scanned from Jacoby (1998)*

30.3. Extension of Coplot

In case the conditioning variable Z is discrete then the plot becomes even simpler. In such a case we can draw scatter plot between X and Y for different values of Z, so the top–most panel *i.e.*, the given panel is not required and the plot looks like a scatterplot matrix but for various values of Z.

30.4. Advantages

(*a*) The plot helps us to understand how the relationship between two variable changes in the presence of another variable.

(*b*) The plot can be used to study the type of relationship and change in the type of relation between the variables.

30.5. Disadvantages

(*a*) Since in the display the scale of the axes, tic marks associated with the axes etc., are concealed so it sometimes becomes difficult to read the values from the figure and also the data can be easily manipulated.

(*b*) As the number variables increases the area available for each panel decreases so only a few variables can be considered for drawing this plot.

(*c*) The coplot is not a very commonly available plot in statistical software.

(*d*) The plot is difficult to draw manually.

30.6. Related Techniques

(*a*) Scatter Plot Matrix;

(*b*) Scatter Diagram; and

(*c*) Trellis Display.

31. CORRELATION PLOT OR CROSS-CORRELATION PLOT

31.1. Definition and Description

Cross-correlation is a measure of the degree of the linear relationship between two data sets. It is similar to autocorrelation except that it compares values in two different data sets insteed of comparing different values within the same data set. Here the correlation between two time series for several lags is computed. They are then plotted in the graph. Thus the plot consists of the lags across the X axis and the corresponding correlations along the Y axis. The range of the Y axis is from −1 to +1. Thus a high correlation (either positive or negative) for a particular lag can be easily detected from the plot.

31.2. Working Data

The data used for this purpose is given in Table 28.1.

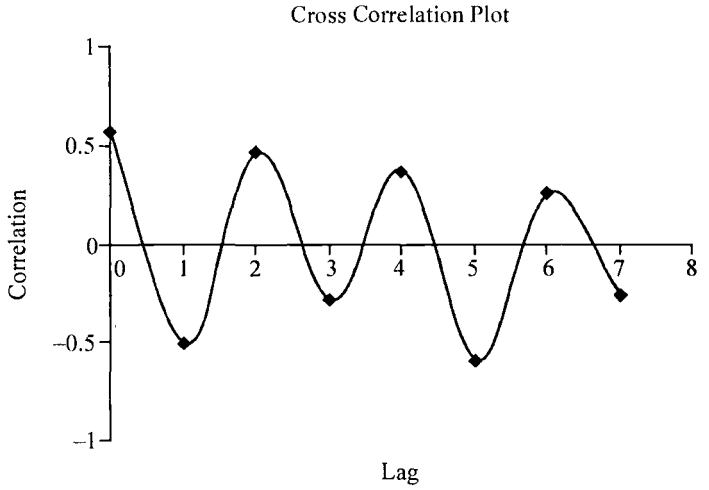


Fig. 31.1. *A Cross-correlation plot for the data in Table 28 1*

The figure shows that at none of the lags the correlation is high. Even at lag 0 or at lag 1 where there is more chance of any relationship the correlation is not very high. Thus we may conclude from the plot that there is no linear relationship between availability of cereals and pulses for any of the lags.

X Axis: In the axis the lags are considered. So the value starts from 1 and may go at least up to k, where $k = \frac{n}{4}$, converted to nearest integer in case k is fractional.

Y Axis: In the vertical axis we consider the correlation coefficients for different lags. So the maximum range of the axes is from −1 to +1.

31.4. Uses

(*a*) The plot is used to study if two time series are independent of each other or not.

(*b*) The plot can be used to understand the dependence if a lagged series has some effect on another time series.

31.5. Related Techniques

(*a*) Auto-correlation Plot;

(*b*) Serial Correlation;

(*c*) Lag Plot; and

(*d*) Partial Autocorrelation Plot.

32. CORRGRAM

32.1. Definition and Description

Often in case of hyper dimensional data (data multiple number of variables) we generally calculate the correlation coefficient of each variable with the other variable. This gives us an idea about the extent of linear relationship between the variables. The correlation coefficients are arranged in the form of a matrix with the diagonal elements equal to one and any other $(i, j)^{th}$ element is the correlation coefficient between X_i and X_j. Thus in the correlation matrix we will have the $(i, j)^{th}$ element is equal to the $(j, i)^{th}$ element. The corrgram is a diagrammatic representation of the correlation matrix. The idea was initiated by Friendly in 2002. If there are p variables then we initially draw p^2 small squares in p rows and p columns. Thus each row as well as each column has p squares. The squares are colored either blue or red, blue color is used for positive correlations and red color is used for negative correlations. The intensity of the color also increases with increase in the magnitude of the correlation coefficient. The squares in the principal diagonal are kept blank. But, here since color is not available so different shades of grey between white to black are used, with intensity of color increasing with increase in the absolute value of correlation coefficient.

32.2. Working Data

The working data used for drawing the corrgram is generated using MS Excel. The table consists of 5 variables each having 10 observations.

Table 32.1: Random data generated in Excel for 5 variables

X_1	X_2	X_3	X_4	X_5
0.79	2.85	6.62	8.03	6.13
1.25	7.92	5.84	3.27	0.71
0.52	0.02	0.45	5.16	5.94
4.96	7.12	8.95	4.14	1.11
3.60	8.59	4.88	6.68	9.67
2.48	7.99	5.01	3.31	8.02
2.09	7.49	0.29	0.12	3.14
1.87	6.75	7.99	4.12	7.73
1.14	8.66	3.28	0.7	2.81
7.57	6.13	8.56	4.41	3.22

Based on the above data we perform the calculations and we get the correlation matrix as follows:

$$\begin{pmatrix} 1 & 0.29 & 0.57 & 0.6 & -0.16 \\ 0.29 & 1 & 0.24 & -0.49 & -0.13 \\ 0.57 & 0.24 & 1 & 0.4 & -0.09 \\ 0.06 & -0.49 & 0.4 & 1 & 0.49 \\ -0.16 & -0.13 & -0.09 & 0.49 & 1 \end{pmatrix}$$

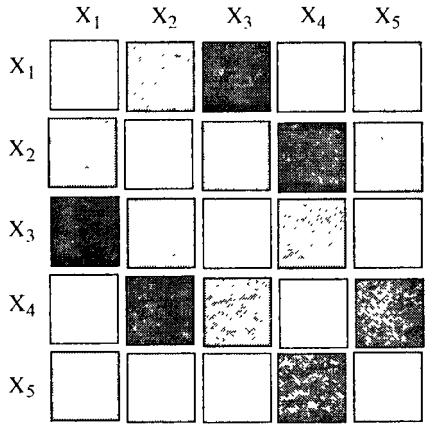


Fig. 32.1. *A Cross-correlation plot for the data in Table 32.1*

32.3. Interpretation of the Plot

Fig. 32.1 shows that at none of the correlations can be considered as high correlation. However the correlation between X_4 and X_2 and that between X_3 and X_1 appears to be in the higher side. The actual corrgram defined in Friendly (2002) is a bit different and has many added features. This is a simpler one and is based on the ideas forwarded by Friendly.

32.4. Advantages

(*a*) The plot gives a quick understanding of the correlation matrix.

(*b*) The variables which have greater extent of linear relation are easier to identify.

(*c*) The interpretation of the plot is as simple as the correlation matrix itself.

32.5. Disadvantages

(*a*) The plot cannot be drawn manually.

(*b*) Small difference between correlation coefficients cannot be identified by visual inspection.

(*c*) The plot is rarely available in any statistical software.

32.6. Related Techniques

1. Correlation Matrix; and
2. Mosaic Plot.

33. CUBES

33.1. Definition and Description

Cubes are three dimensional diagrams. They cannot be drawn properly in paper as the paper is a two dimensional surface. Cubes have equal length, breadth and height so these diagrams are also called as volume diagrams. Here the cube root of the response variable is initially calculated. Then cubes with sides proportional to the values of the cube roots are drawn.

33.2. Working Data

The data used for the plot is a hypothetical figure representing sale of a particular product launched by a company corresponding to the first four months are given below.

Table 33.1: Sale of a particular product in '000.

Month	*Sale (in '000)*	*Cube Roots*
1	12	2.29
2	70	4.12
3	150	5.31
4	270	6.47

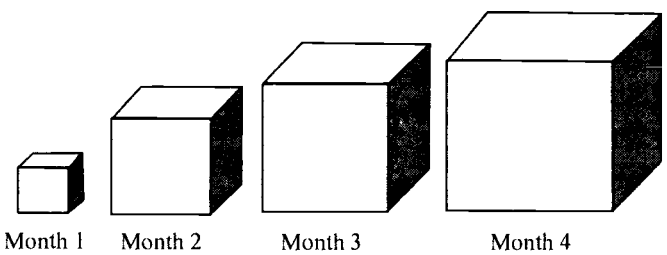


Fig. 33.1. *Cubes corresponding to data in Table 33.1*

33.3. Axes

The diagram does not require the help of any axes.

33.4. Uses

(*a*) The plot is used for the representation of values of one variable only.

(*b*) The data now provides the option for three dimensional viewing of a set of data.

33.5. Related Techniques

(*a*) Sphere;
(*b*) Pyramid;
(*c*) Cone; and
(*d*) Cylinder.

34. CUSUM CONTROL CHART

34.1. Definition and Description

In a production process when the quality of the product can be measured the Cumulative Sum Control Charts or CUSUM Charts can be used to detect small shifts in the mean of the process. In order to construct the CUSUM chart we first select randomly k samples each one of size n and compute the mean of each sample, *i.e.*, $\bar{x}_i$. Next we compute $S_m = \sum_{i=1}^{m}(\bar{x}_i - \hat{\mu}_0)$. Then the (CUSUM) chart is formed by plotting S_m against m, where m denotes the sample number. Here $\hat{\mu}_0$ represents the estimate of the in control mean. As long as the process remains within control the plotted points centers around 0. However, if the process mean shift upward, the cusum points will start drifting upwards, and will drift downwards if the process mean decreases.

34.2. Working Data

The data for this purpose is the life in hours of cells after complete charging obtained from a local manufacturer of cells. The data consists of 15 samples each of size 5. The company wants to produce cells with an average life of 27 hours. Thus $\hat{\mu}_0 = = 27$

Table 34.1: Life in hours of cells after full charging

Sample No.	*Life of Cells in Hours*					*Mean* $(\bar{x}_i)$	S_m
1	30.5	34.6	20.2	29.0	30.7	26.56	–0.44
2	23.9	22.1	38.7	39.8	27.5	28.29	0.85
3	31.1	23.5	31.4	27.3	39.6	28.17	2.02
4	24.3	39.4	30.9	21.9	25.2	26.54	1.56
5	20.7	20.6	37.7	26.3	32.3	25.79	0.35
6	36.6	27.1	20.1	33.8	29.9	27.33	0.68
7	37.7	27.9	20.2	25.7	36.2	27.53	1.21
8	26.7	38.4	22.4	29.4	24.1	26.17	0.39
9	29.7	20.3	30.6	35.2	36.1	27.96	1.34
10	22.8	33.3	32.3	33.0	27.2	26.52	0.86
11	25.9	22.4	29.3	28.2	21.6	22.52	–3.62
12	30.1	31.6	24.4	27.1	27.7	24.69	–5.93
13	22.1	26.6	38.3	33.2	29.6	27.67	–5.26

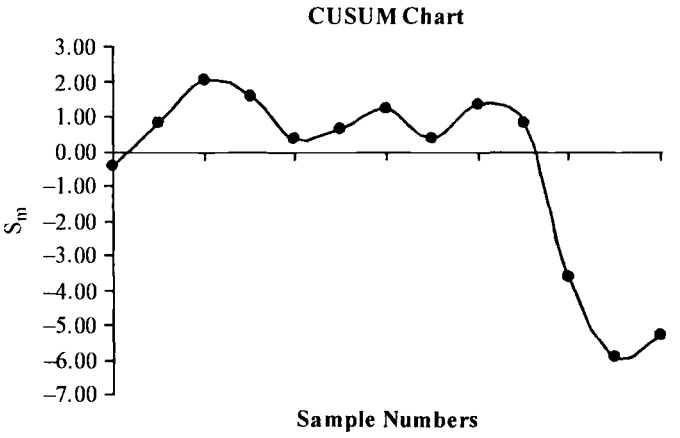


Fig. 34.1. *The CUSUM chart drawn for the data provided in Table 34.1*

Thus, we see that the shift is towards the negative direction with increase in the sample number which implies that there is a shift in the mean life of the cells.

34.3. Axes

X Axis: The axis is used to represent the sample numbers.

Y Axis: This axis is used to represent the values of S_m.

34.4. Uses

(*a*) The plot is more sensitive compared to the $\bar{x}$ chart, R chart, etc.

(*b*) The plot can be used to analyse the shift in process mean.

(*c*) The chart can be used even if the number of samples is small in number.

34.5. Related Techniques

(*a*) Process Control;

(*b*) Control Chart;

(*c*) *s* Chart;

(*d*) $\bar{X}$ Chart;

(*e*) *p* Chart; and

(*f*) EWMA Chart.

35. DUANE PLOT

35.1. Definition and Description

The plot is related to reliability statistics. If we have the failure times of a number of k components T_i, $(i = 1, 2, \ldots, k)$ then the Duane plot is the plot between $\frac{T_i}{i}$ versus T_i in a log–log scale. The Duane plot is a plot of cumulative operating time against the cumulative failure rate on double logarithmic axes. The plot is used to detect if the cumulative failure time of a system follows Non–Homogeneous Poisson Process (NHPP), *i.e.*, of the form $f(t) = \alpha t^{-\beta}$ or not. The plot was developed by J.T. Duane in 1964. In case the cumulative failure time follows NHPP then the plotted points fall in a straight line. In practice before constructing a Duane plot some other plots for trend determination is generally applied.

35.2. Working Data

A system is subjected to operation for 2000 operational hours and in that period there were 10 failures. The observed failure times were: 505, 540, 553, 675, 889, 1212, 1247, 1395, 1799 and 1978 hours, with the test ending at 2000 hours.

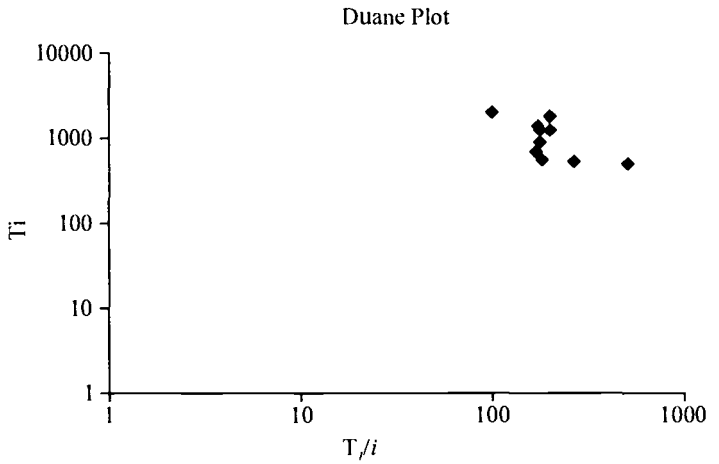


Fig. 35.1. *Duane Plot corresponding to the data stated above.*

Since the points are not in a straight line so we may conclude that the cumulative failure time of a system does not follows Non–Homogeneous Poisson Process (NHPP).

35.3. Axes

X Axis: The axis is used to represent $\frac{T_i}{i}$, where T_i is the time of failure of the component i^{th} component.

Y Axis: Along this axes T_i the time of failure of the components are considered.

35.4. Uses

(*a*) The plot is used to detect if the cumulative failure time of a system follows Non-Homogeneous Poisson Process (NHPP) *i.e.*, of the form $f(t) = \alpha t^{-\beta}$ or not.

(*b*) If the points fall in an approximate straight line then the slope of the fitted line to the points can be used as an estimate of β

35.5. Related Techniques

(*a*) MTBF;

(*b*) Reliability Improvement Test; and

(*c*) Inter Arrival Time Plot.

36. DENDROGRAM

36.1. Definition and Description

Dendrogram is a graphical technique which helps in the clustering of multivariate observations. Let us consider n variables each having N observations. At the initial stage each observation is considered to form a separate cluster. So initially for N observations we consider N clusters. Now, for each pair of multivariate observation we compute the Euclidean distance. Let $\underline{x} = (x_1, x_2, ..., x_n)$ and $\underline{y} = (y_1, y_2, ..., y_n)$ be two multivariate observations then the Euclidian distance is given by,

$$d(x,y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + ... + (x_n - y_n)^2}$$

On computing the Euclidean distance for all pairs of observations, a distance matrix N × N is constructed. The pair of observation which has the minimum Euclidean distance is considered in the same cluster. Without any loss of generality let us assume that observation 1 and 2 has the minimum Euclidean distance. Let us name the newly formed clusters as (1, 2). Update the entries in the distance matrix by deleting the rows and columns corresponding to cluster 1 and 2, *i.e.*, the first and second rows and columns in this case. In the next step we compute the Euclidean distance between the cluster (1, 2) and the remaining clusters and accordingly adding a row and column giving the distance between cluster (1, 2) and the remaining clusters. This process is repeated $N - 1$ times. This technique is called as single linkage.

The results of the single linkage are graphically displayed in the form of a tree diagram called as the dendrogram. The branches of the dendrogram is represents the different clusters. Looking at the point of connection of the branches and the formation of the nodes one can make out the stage at which the clustering took place.

36.2. Working Data

The data set is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups.

Table 36.1: Protein consumption in European countries

	Red meat	*White meat*	*Eggs*	*Milk*	*Fish*	*Cereals*	*Starchy food*	*Pulses*	*Fruits veg.*
Albania	10.1	1.4	0.5	8.9	0.2	42.3	0.6	5.5	1.7
Austria	8.9	14	4.3	19 9	2.1	28	3.6	1.3	4.3
Belgium	13.5	9.3	4.1	17.5	4.5	26.6	5.7	2.1	4
Bulgaria	7.8	6	1.6	8.3	1.2	56.7	1.1	3.7	4.2
Czechoslovakia	9.7	11.4	2.8	12.5	2	34.3	5	1.1	4
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4

(contd...)

	Red meat	*White meat*	*Eggs*	*Milk*	*Fish*	*Cereals*	*Starchy food*	*Pulses*	*Fruits veg.*
Denmark	10.6	10.8	3.7	25	9.9	21.9	4.8	0.7	2.4
East Germany	8.4	11.6	3.7	11.1	5.4	24.6	6.5	0.8	3.6
Finland	9.5	4.9	2.7	33.7	5.8	26.3	5.1	1	1.4
France	18	9.9	3.3	19.5	5.7	28.1	4.8	2.4	6.5
Greece	10.2	3	2.8	17.6	5.9	45.7	2.2	7.8	6.5
Hungery	5.3	12.4	2.9	9.7	0.3	40.1	4	5.4	4.2
Ireland	13.9	10	4.7	25.8	2.2	24	6.2	1.6	2.9
Italy	9	5.1	2.9	13.7	3.4	36.8	2.1	4.3	6.7
Netherland	9.5	13.6	3.6	23.4	2.5	22.4	4.2	1.8	3.7
Norway	9.4	4.7	2.7	23.3	9.7	23	4.6	1.6	2.7
Poland	6.9	10.2	2.7	19.3	3	36.1	5.9	2	6.6
Portugal	6.2	3.7	1.1	4.9	14.2	27	5.9	4.7	7.9
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Romania	6.2	6.3	1.5	11.1	1	49.6	3.1	5.3	2.8
Spain	7.1	3.4	3.1	8.6	7	29.2	5.7	5.9	7.2
Sweden	9.9	7.8	3.5	24.7	7.5	19.5	3.7	1.4	2
Switzarland	13.1	10.1	3 1	23.8	2.3	25.6	2.8	2.4	4.9
UK	17.4	5.7	4.7	20.6	4.3	24.3	4.7	3.4	3.3
USSR	9.3	4.6	2.1	16.6	3	43.6	6.4	3.4	2.9
West Germany	11.4	12.5	4.1	18.8	3.4	18.6	5.2	1.5	3.8

A SINGLE LINKAGE DENDROGRAM

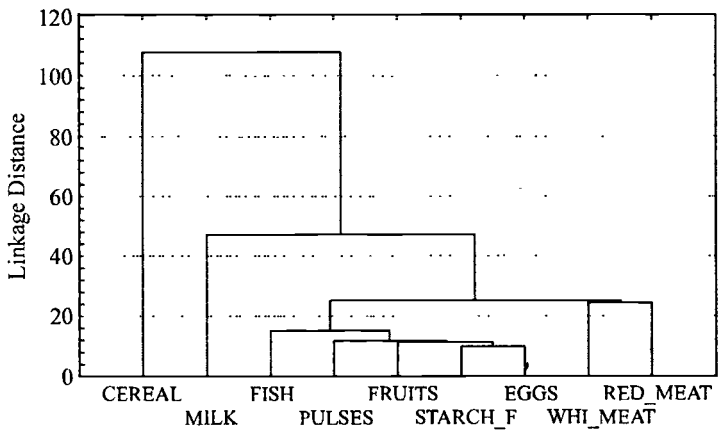


Fig. 36.1. *A Dendrogram for data provided in Table 36.1*

From the plot we can say that if we want to divide the food items into two clusters only than we have 'Cereals' in one cluster and all others in the other clusters. In case of three clusters it can be seen from the diagram that we have 'Cereals' in one cluster, 'Milk' in another and all other food items in other cluster, and so on.

36.3. Axes

X Axis: It can represent the different categories. Here different types of food items are considered along the X–axis.

Y Axis: The vertical axis we consider the values of the Euclidian distance.

36.4. Uses

(*a*) To visualize the single linkage clustering.

(*b*) To visualize the Euclidian distance between any two observations.

(*c*) To divide the data into any number of clusters.

(*d*) Has extensive use in social sciences, ecological science, medical science.

36.5. Some Related Techniques

(*a*) Hierarchical Clustering Methods;

(*b*) Complete Linkage Clustering; and

(*c*) Classification.

For further reading related to the plot and about calculation of the Euclidean Matrix the following reference may be consulted:

- Johnson, R. A. and Wichern, D. W. (1992) *Applied Multivariate Statistical Analysis*, 3rd Edition, Prentice-Hall International, N. J, USA.

37. DETRENDED PROBABILITY PLOT

36.1. Definition and Description

Whenever we express a variable with that of the other in terms of a mathematical relation then the most general form in Statistics is given by

$$y = f(x) + e,$$

where, the variable 'e' is called as the error component associated with the model. In case of absence of the term 'e' from the model the relation would be a deterministic one, which is definitely an ideal situation. But such an ideal situation rarely exists in reality, whenever we deal with a practical data set. If we fit a good model for a data set then there will be no connection between e and $f(x)$, *i.e.*, in other words, between 'e' and 'x'. In such a case we may say that we have found an appropriate model involving 'x' that can be used for predicting 'y'. On the other hand, if the fit is not a good one, then there may be other variables which may play a role in predicting y the influence of all such variables will be contained in the error component 'e'. Generally, we will not know the values of 'e' for all the given values of 'x'. Thus, we can estimate the coefficients involved in the function $f(x)$ and accordingly estimate the values of 'y' for a given 'x', which is denoted by '$\hat{y}$'. The difference 'y' and '$\hat{y}$', *i.e.*, '$y - \hat{y}$' is called as the residual. The significance of the residual is to measure the discrepancy of an observation from some reference value that is been specified by the fitted model. In case of a good fit these residuals are independent and identically distributed and generally follow some fixed probability law. In most cases it is assumed that the residuals follow normal probability law. Thus we calculate the residuals for each of the observations and accordingly draw the normal probability plot. Such a normal probability plot is called as detrended probability plot.

(Readers are requested to read '72. Normal Probability Plot' for details about the plotting technique)

37.2. Working Data

The data for this purpose is some hypothetical data for two variables X and Y.

Table 37.1: A Hypothetical data showing the values of X and Y along with the trend values from a fitted line and corresponding residuals

X	Y	$\hat{Y} = 3.936X + 16.472$	*Residuals(e)*
1	13.06	20.4081	–7.3481
2	18.44	24.3442	–5 9042
4	28.96	32.2164	–3.2564
6	39.16	40.0886	–0.9286
8	49.04	47.9608	1.0792

(contd...)

X	Y	$\hat{Y} = 3.936X + 16.472$	*Residuals(e)*
9	53.86	51.8969	1.9631
12	67.84	63.7052	4.1348
14	76.76	71.5774	5.1826
16	85.36	79.4496	5.9104
23	112.94	107.0023	5.9377
26	123.56	118.8106	4.7494
28	130.24	126.6828	3.5572
34	148.36	150.2994	−1.9394
36	153.76	158.1716	−4.4116
39	161.26	169.9799	−8.7199

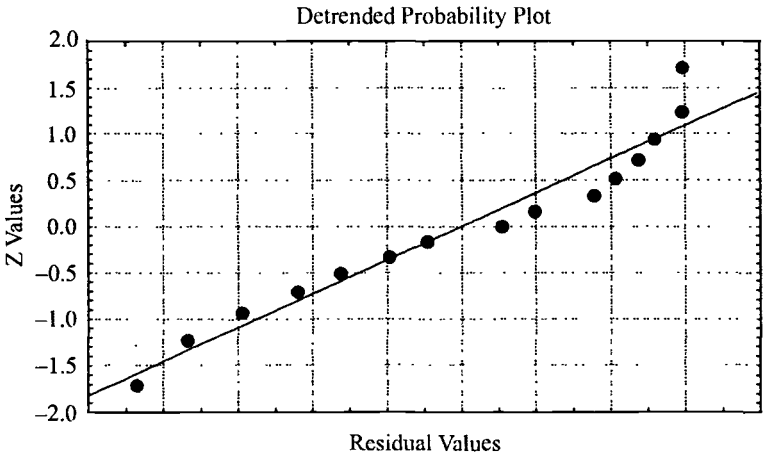


Fig. 37.1. *The detrended probability plot based on data in Table 37.1*

Thus, we see that in the residual values are not normally distributed as the points falls away from the hypothetical line.

37.3. Axes

X Axis: The axis is used to represent the residuals.

Y Axis: This axis is used to represent the Z values corresponding to each value of the residual.

37.4. Advantages

(*a*) The plot provides a quick check for the normality of the residuals from a model.

(*b*) The normality of the residuals can be performed both for large as well as small samples.

(*c*) The detrended probability plot can be used for a large variety of models.

37.5. Disadvantage

The most striking problem with the plot is that different people reach to different conclusions regarding the normality of the residuals. Since, no proper confidence bands are developed around the straight line obtained from the plotting of expected frequency from the hypothetical distribution, so accessing the goodness of fit sometimes, becomes difficult even for experts.

37.6. Related Techniques

(*a*) Probability Plot;
(*b*) Normal Probability Plot;
(*c*) Chi-Square Test;
(*d*) Anderson-Darling test; and
(*e*) Wilks Shapiro test.

38. DEVIATION PLOT

38.1. Definition and Description

There are different types of deviation plots and different software has used the name for plots completely different from each other. The deviation plot that we are going to discuss may also be termed as the "bar deviation plot". In this plot the X axis generally comprise of various categories and the values of the response variable are plotted along the Y axis. The individual data points are then represented by vertical bars. However, the bars connect the data points to a user-selectable baseline (parallel to X-axis). If the baseline value is different than the plot's Y-axis minimum, then individual bars will extend either up or down, depending on the direction of the "deviation" of individual data points from the baseline. Ideally if one considers the Y-axis as the base line then the plot will look almost like a bar diagram. However it may be recommended that the average value of Y may be used to draw a base line and accordingly the values above and below the averages can be identified.

38.2. Working Data

The data used for the plot is obtained from a survey conducted on the internet users of a town. 80 people were surveyed and their favorite Internet Café was noted. Accordingly the following data was obtained.

Table 38.1: Favorite internet cafe of the town

Name of the Café	*No. of Respondents*
Cyber Max	22
A to Z	20
Cyber World	17
Chit-o-Chat	10
Impact Graphics	7
Net Zone	4
Total	80

38.3. Axes

X Axis: It represents the various categories corresponding to which the values are provided. Here the various Internet Café is considered as the categories.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took the number of respondents along the Y axis.

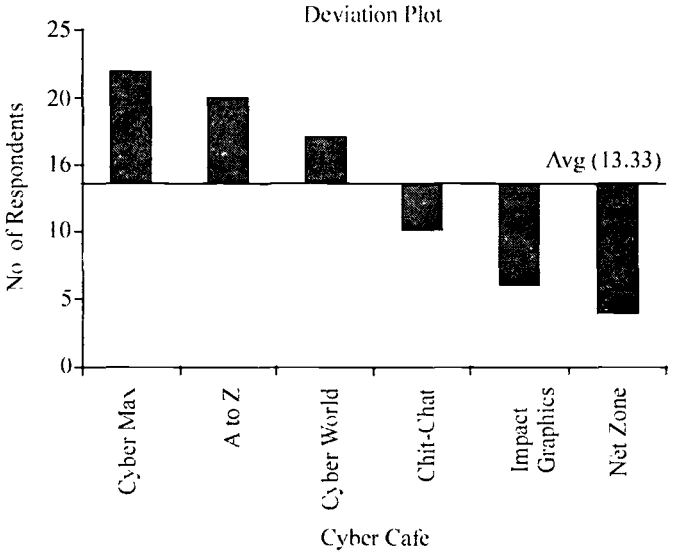


Fig. 38.1. *A Deviation Plot for the data in Table 38.1*

38.4. Uses

(*a*) The diagram can be used for data representation especially in cases where the magnitudes of one or more variables under different categories (or classes) are to be compared.

(*b*) It can also be used for the visualization of the mean of a set of observations.

(*c*) To understand which categories are performing below the average and which are above it.

38.5. Related Techniques

(*a*) Bar Diagram;

(*b*) Residual Bar Plot; and

(*c*) Column Plot.

39. DIGRAPH

39.1. Definition and Description

This plot is used for the representation of a transition probability matrix which is often encountered in the study of Markov chain. A transition probability matrix is a matrix consisting of transition probabilities. The transition probability p_{ij} is defined as,

$$p_{ij} = P[X_n = j \mid X_{n-1} = i]$$

where X_n is a stochastic process.

For drawing the graph the states are first noted down and encircled with a finite distance from one another. If the state i is approachable from state j then the two states are connected by curved lines. If $j > i$, then the curves are drawn above the circles otherwise they are drawn below the circles. The curves are marked with arrows in order to show the direction of movement.

39.2. Working Data

The data for this purpose is a transition probability matrix taken from Medhi (1994). There are 4 states and the matrix is as:

$$P = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{pmatrix} 1/6 & 1/3 & 1/2 & 0 \\ 1/2 & 1/2 & 0 & 0 \\ 1/6 & 1/3 & 1/2 & 0 \\ 0 & 1/6 & 1/3 & 1/2 \end{pmatrix} \end{array}$$

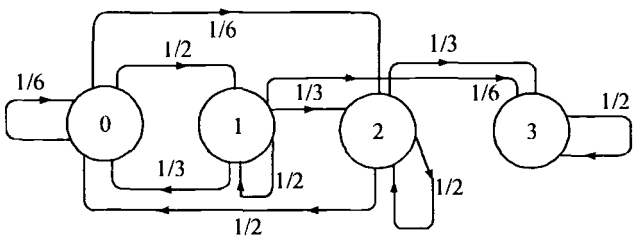


Fig. 39.1. *A Digraph for transition probability matrix provided above*

39.3. Axes

No axes are required to draw the graph.

39.4. Uses

(*a*) To visualize the transition probability matrix.

(*b*) To find the most communicative state and the least communicative one from the graph.

39.5. Some Related Techniques

(*a*) Transition Probabilities;

(*b*) Markov Chain; and

(*c*) Transition Probability Matrix.

40. DOE MEAN PLOT

40.1. Definition and Description

The DOE mean plot may be considered as a graphical counter part of the analysis of variance technique. The technique can be applied for analyzing data from a designed experiment, with respect to important factors, where the factors are at two or more levels. The plot shows mean values for the two or more levels of each factor plotted for the different factors. The factors are taken along the X axis and the mean responses of each factor under each levels are taken along the Y axis. Here the grand mean of all the observations are computed first and is plotted by a straight line parallel to X axis. The means of response of each factor for different levels are then plotted. The means at different levels for a single factor are then connected by a straight line. The X-coordinate of the levels for a particular factor are kept slightly different from each other so that the lines are not perpendicular to the line representing the grand mean.

40.2. Working Data

The working data is taken from Barlett (1936) which gives the values of number of surviving latherjackets for different blocks and emulsions. The data is provided in Table 13.1

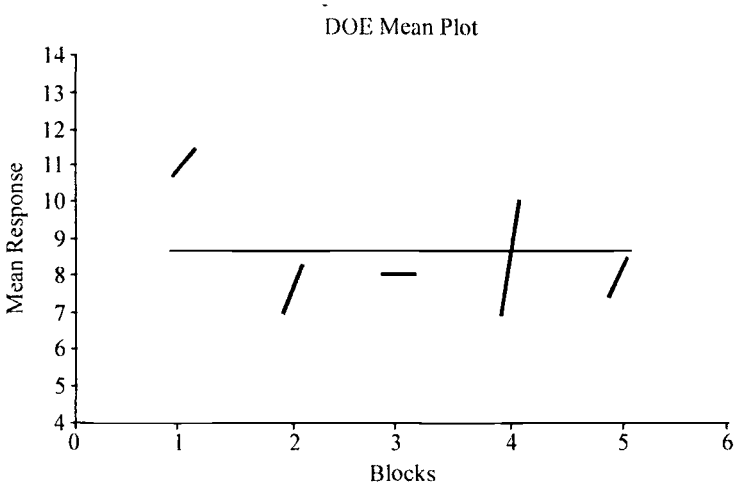


Fig. 40.1. *DOE Mean plot for data in Table 13.1*

The plot shows that the mean is maximum for both the levels in case of Block 1 and Block 2 is probably the worst performing factor. In all the factors the performance of Level 2 is better than Level 1. The variation between the two levels is maximum in case of Block 4 and minimum in case of Block 3.

40.3. Axes

X Axis: The different factors are taken along the X axis, here the Blocks are taken along the axis.

Y Axis: The mean of the response variable for each factor at each level is considered along the axis.

40.4. Advantages

(*a*) Gives us an idea about which factor may be considered as the most important one, and helps in ranking the factors based on its relative importance.

(*b*) Helps in the visualization of difference between two levels of the same factor.

40.5. Disadvantages

(*a*) Cannot study if the difference between means of the response variables differs significantly at a given level of significance. This restricts the use of the plot to inferential problems.

(*b*) Most statistical packages do not provide an option to draw the plot.

40.6. Related Techniques

(*a*) DOE Scatter plot;

(*b*) Analysis of Variance; and

(*c*) Factorial Experiment.

41. DOE SCATTER PLOT

41.1. Definition and Description

The DOE scatter plot may be considered as a graphical counter part of the analysis of variance technique. The technique can be applied for analyzing data from a designed experiment, with respect to different factors, where the factors are at two or more levels. Here the factors are taken along the X axis which are been separated by a distance. The factors may be named or some identifying number can be used. The responses obtained under each factor are plotted along the Y axis. The responses at different levels are been symbolized differently which helps in the identification of the levels. Also the mean value of each treatment is shown as a solid black bar in the diagram amongst the cluster of points corresponding to each treatment. This makes the means readily visible. In order to avoid over plotting of the plotted points, a random number between [–0.2, 0.2] is generated and is added to the group identifier variable before plotting. This disturbs the horizontal alignment of the points belonging to a particular factor and hence acts a solution to the problem of overplotting.

41.2. Working Data

The working data is taken from Barlett (1936) which gives the values of number of surviving latherjackets for different controls and emulsions. The data is provided in Table 13.1

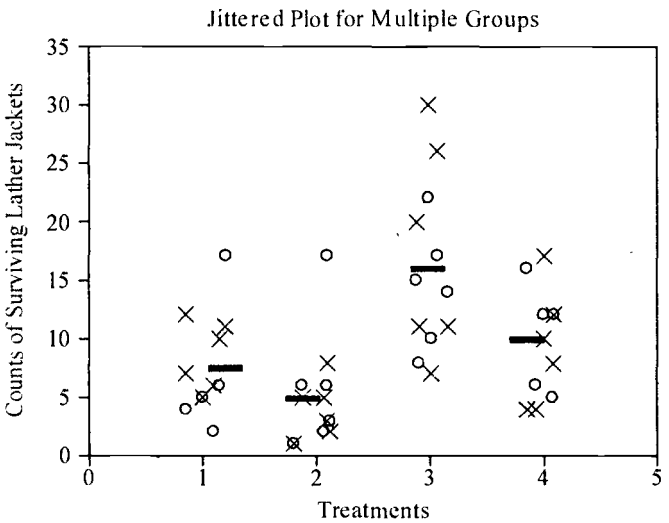


Fig. 41.1. *DOE scatter plot for data in Table 13.1*

The plot shows that the observations under treatment 3 shows more variation compared to the others and also the mean appears to be more in case of blocks receiving the third treatment. Level 1 (circle) under the first two treatments produces outliers. The two levels do not produce significant difference in the response variable across the various treatments.

41.3. Axes

X Axis: The different factors are taken along the X axis, here the Emulsions are taken along the axis.

Y Axis: The values of the response variable for each factor at each level are considered along the axis. Here counts of surviving lather jackets are considered along the Y axis.

41.4. Advantages

(*a*) Gives us an idea about which factor may be considered as the most important one, and helps in ranking the factors based on its relative importance.

(*b*) Helps in the visual comparison between the means of different factors.

(*c*) The variability in the values of the response variables under different factors can also be considered.

(*d*) The plot can also be used for the identification of outliers.

41.5. Disadvantages

(*a*) Cannot study if the difference between means of the response variables under different factors differs significantly at a given level of significance. This restricts the use of the plot to inferential problems.

(*b*) Most statistical packages does not provide an option to draw the plot.

41.6. Related Techniques

(*a*) DOE Mean Plot;

(*b*) DOE Standard Deviation Plot;

(*c*) Analysis of Variance; and

(*d*) Factorial Experiment.

42. DOUBLE Y-AXIS PLOT

42.1. Definition and Description

This plot is also used for the representation of three numerical variables. When there is one independent variable and two other variables that are dependent on the independent variable. The two dependent variables can have different scale and/or units of measurement and so cannot be plotted in a common vertical scale.

In this plot there are two Y-axis and hence the name. Both the axes are parallel to each other but are perpendicular to the X-axis. The independent variable (time in case of a time series plot) is taken along the X-axis. One of the dependent variable is plotted and accordingly they are connected by lines or by free hand smooth curve. The other dependent variable is plotted corresponding to the other Y axis which is scaled differently and placed along the right most side of the graph. The plotted points are then connected by straight lines or free hand smooth curve. The two curves thus obtained are colored differently so that it can be easily referred to the respective axis. One can color the Y-axes in the same color as its corresponding curve so that it is easier to relate.

42.2. Working Data

For drawing a Double Y-axis graph we need to have three variables two of which are the response variable and the other one being the independent variable. The data used for drawing such a graph is shown in Table 42.1. The data comprises of the census figures of India related to population density and literacy rate taken from census report of 2001.

Table 42.1: Literacy rate and population density of India for the various census years

Census Years	*Literacy Rate (%)*	*Density (No. of people per sq. km)*
1901	5.35	77
1911	5.92	82
1921	7.16	81
1931	9.5	90
1941	16.1	103
1951	16.67	117
1961	24.02	142
1971	29.45	177
1981	36.23	216
1991	42.84	267
2001	55.3	324

Source: Provisional Statistics, Census — 2001

From the table we find that both the factors i.e. population density as well as literacy rate is dependent on time, this make the data ideal for visualization using double *Y*-axes plot.

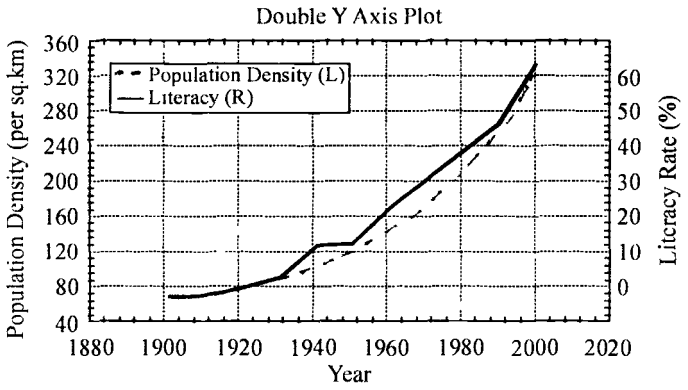


Fig. 42.1. *Double Y-axis plot used for representing data in Table 42.1*

42.3. Axes

For the plot we have three (3) axes viz. *X* axis, first *Y* axis and second *Y* axis.

X-Axis: It can represents the values of the variable that we suspect may have a relation to the response variable for time series plot time is taken along this axis.

First Y-Axis: The first vertical axis consists of one of the variables which we consider as the response variable.

Second Y-Axis: The second vertical axis kept parallel to the first vertical axis and used to represent the second response variable.

42.4. Advantage

(*a*) The basic purpose of the plot is to represent three variables in two dimensions.

(*b*) The same graph can be used to represent two response variables that differs in their units.

(*c*) This graph should be complemented for saving the space as otherwise one should have drawn two separate graphs for the same representation.

42.5. Disadvantage

Since the scales of the two vertical axes are different and also is their units, so they should not be used for comparison in absolute terms. Even it is not safe to compare the rate of change of the two response variable as it is easier to manipulate the scales to produce a different picture.

42.6. Related Techniques

(*a*) Bubble Plot;

(*b*) Categorical Scatter Plot; and

(*c*) Glyph Plot.

43. DROUGHNUT CHART

43.1. Definition and Description

The chart can be considered as the multivariate extension of a pie chart. Such diagrams are used representing how different values of a response variable is sub-dividing into components under different categories. Here we draw some concentric circles with increasing radius. In case of n value of the response variable $n + 1$ such concentric circles are drawn. Then starting from the center of the circle and moving out radially, the area bounded by two consecutive concentric circles are subdivided into several sectors like that of a pie diagram using the components of the response variable under different categories. Thus, the first two concentric circles are used to represent the components of the first value of the response variable under different categories. The different sectors between two concentric circles are coloured or shaded differently. The legend accompanying the graph can be used for indexing the various shades/colours.

43.2. Working Data

The data used for this plot is provided in Table 10.1.

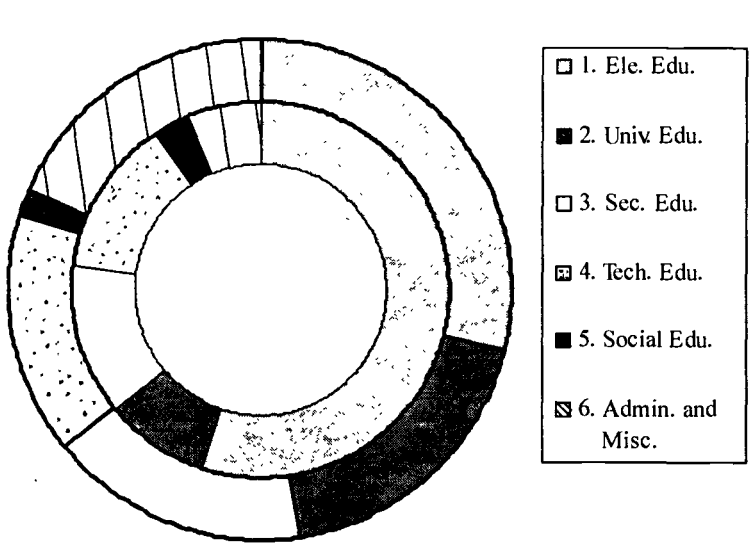


Fig. 43.1. *Doughnut Chart based on data in Table 10.1.*

43.3. Axes

The chart does not require any axes.

43.4. Advantages

(*a*) The plot is especially used in cases when the various components of the response variable are to be compared.

(*b*) The chart can also be used to see how the relative share of the different components varies across different values of the response variable.

(*c*) The graph can be drawn even without the use of color by using the shading option.

(*d*) The chart is available in MS Excel by default.

43.5. Disadvantages

(*a*) The chart can be used for comparison of the share of the components only. However, the actual values of the components cannot be compared from the data.

(*b*) The chart is difficult to understand compared to a sub-divided bar diagram that can be used for the same purpose.

(*c*) If not drawn with the help of software, the chart involves a lot of calculations.

43.5. Related Techniques

(*a*) Bar Diagram;

(*b*) Pie Chart;

(*c*) Sub-divided Bar Diagram; and

(*d*) Pie Icon Plot.

44. DUBEY PLOT

44.1. Definition and Description

Dubey (1966) and Rao (1971) working separately derived a simple technique that can be used as graphical tests for discrete distributions. The technique derived by them, uses the ratio of the probability mass function for consecutive values of the random variable. Dubey (1966) showed that the ratio of $P[X = x + 1] : P[X = x]$ takes a linear form for binomial, Poisson and geometric distribution. This can be easily extended to some other discrete distributions.

In case of binomial distribution, we have

$$P[X=x] = {}^nC_x p^x (1-p)^{n-x},\ x = 0, 1, 2, \ldots, n$$

$$\Rightarrow \quad \frac{P[X=x+1]}{P[X=x]} = {}^nC_{x+1} p^{x+1}(1-p)^{n-x-1} \Big/ {}^nC_x p^x (1-p)^{n-x} = \frac{n-x}{x+1} \times \frac{p}{1-p}$$

Thus,
$$\frac{P[X=x+1]}{P[X=x]} = -\frac{p}{1-p} + \left(\frac{(n+1)p}{q}\right)\frac{1}{x+1} \qquad \ldots(44.1)$$

Now, if we consider the equation (44.1) then we find that the expression

$P[X=x+1] : P[X=x]$ is a linear function with respect to $z = \frac{1}{x+1}$.

In case of Poisson distribution we have,

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!},\ \lambda > 0, x = 0, 1, 2, \ldots$$

$$\Rightarrow \quad \frac{P[X=x+1]}{P[X=x]} = \frac{e^{-\lambda}\lambda^{x+1}}{(x+1)!} \Big/ \frac{e^{-\lambda}\lambda^x}{(x)!} = \frac{\lambda}{x+1}$$

Thus,
$$\frac{P[X=x+1]}{P[X=x]} = \lambda\left(\frac{1}{x+1}\right) \qquad \ldots(44.2)$$

Thus, for a Poisson Distribution also we have $P[X = x + 1] : P[X = x]$ is a linear function with respect to $z = \frac{1}{x+1}$.

However, in case of geometric distribution we have,

$$P(X=x) = q^x p,\ \ 0 < p < 1, x = 0, 1, 2, \ldots$$

$$\Rightarrow \quad \frac{P[X=x+1]}{P[X=x]} = \frac{q^{x+1}p}{q^x p} = q \qquad \ldots(44.3)$$

Thus, the ratio provides a constant function that can also be used to estimate the value of the probability of failure *i.e.*, q.

44.2. Working Data

The data used for drawing this plot is taken from Kemp and Kemp (1991). 12 die were rolled together for 26306 number of times, and accordingly number of success were noted. Here a '5' or a '6' is denoted as success.

Table 44.1: Number of success in 26306 throws of 12 dice

No. of Success	0	1	2	3	4	5	6	7	8	9	10	11	12
Frequency	185	1149	3265	5475	6114	5194	3067	1331	403	105	14	4	0

Before drawing the plot it is essential to convert the data by performing some calculations. The calculations are shown in Table 44.2 below.

Table 44.2: Calculations for Dubey plot based on the data in Table 44.1

No. of Success (X)	*Frequency* (f)	$P[X = x] = f/N$	$\frac{P[X = x+1]}{P[X = x]}$	$\frac{1}{x+1}$
0	185	0.007033	6.210811	1
1	1149	0.043678	2.841601	0.5
2	3265	0.124116	1.676876	0.333333
3	5475	0.208127	1.116712	0.25
4	6114	0.232418	0.849526	0.2
5	5194	0.197445	0.590489	0.166667
6	3067	0.116589	0.433975	0.142857
7	1331	0.050597	0.30278	0.125
8	403	0.01532	0.260546	0.111111
9	105	0.003991	0.133333	0 1
10	14	0.000532	0.285714	0.090909
11	4	0.000152	0	0.083333
12	0	0		0.076923
	$N = 26306$			

44.3. Axes

X Axis: In this plot along the X-axis we have taken the values of $\frac{1}{x+1}$, where x represents the number of success.

Y Axis: The values of $\frac{P[X = x+1]}{P[X = x]}$ are computed from the data and accordingly the values are plotted along the vertical axis.

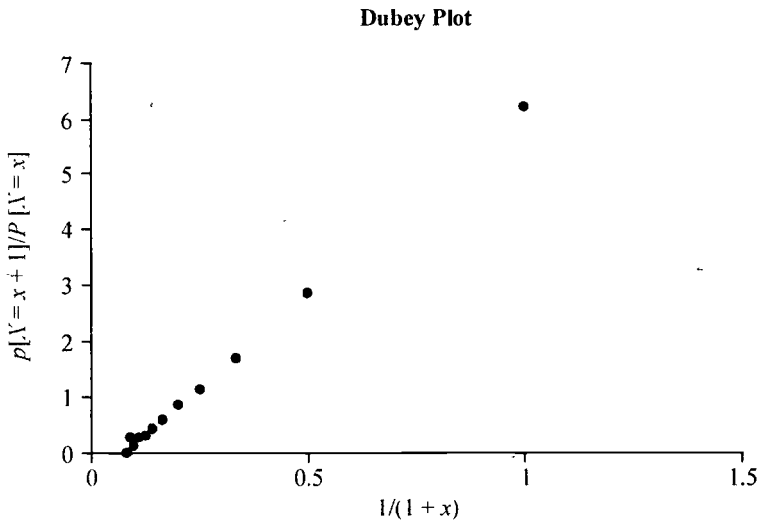


Fig. 44.1. *A Dubey plot to the data in Table 44.1*

From the plot it appears that the points are in a straight line. Thus the data appears to be a good fit either to a Poisson distribution or to a binomial distribution.

44.4. Advantages

(*a*) Dubey Plot provides a quick check to the goodness of fit.

(*b*) Probability plot can be used for goodness of fit of binomial, Poisson and geometric distribution

(*c*) The intercept and slope of the straight line, which the plotted points approximately follow, can be used for the estimation of parameters of the distribution.

44.5. Disadvantages

(*a*) The most striking problem with the plot is that different people reach to different conclusions regarding the goodness of fit test. Since, no proper confidence bands are developed around the straight line obtained from the plotting of expected frequency from the hypothetical distribution, so accessing the goodness of fit sometimes, becomes difficult even for experts.

(*b*) The plot cannot be used to diagnose the distribution. As in this case we cannot find out whether the data fits the binomial distribution or the Poisson distribution.

44.6. Related Techniques

(*a*) Chi-square test for the goodness of fit;

(*b*) Ord Plot;

(*c*) Poissonness Plot; and

(*d*) Binomialness Plot.

45. EMPIRICAL DISTRIBUTION FUNCTION PLOT

45.1. Definition and Description

Let $(y_i, i=1,2,...,n)$ be a random sample of size 'n' on Y, and let $F(y)$ be the cumulative distribution function (CDF) of the random variable. Also, let $(y_{(i)}, i=1,2,...,n)$ denotes the corresponding order statistics. The empirical CDF is given by,

$$F_n(y) = (\text{no. of observations} \leq y)/n$$

Thus this function will be a step function lying between 0 and 1, with jumps of size $1/n$ at each observation. This empirical distribution function popularly called as the EDF, is used as the non–parametric estimate of the actual CDF. The values of $F_n(y)$ are plotted corresponding to the values of y. The cumulative probabilities are computed based on the theoretical distribution that is to be fitted. If the parameters of the theoretical distribution are not known using the maximum likelihood estimator and the parameters can be obtained and accordingly the corresponding CDF curve is drawn. In case the data is a good fit to the corresponding theoretical distribution then the CDF curve and the EDF step polygon will lie in close proximity to each other otherwise they will be separated apart.

45.2. Working Data and Calculations

The data collected for the purpose is taken from Anderson (1958) which pertains the head length of a group of individuals. To check if the observations can be considered to have come from normal distribution we draw the EDF plot and compare it with the corresponding CDF. If the parameters of the normal distribution (μ, σ^2) are unknown we can estimate the parameters using their corresponding maximum likelihood estimators *i.e.*,

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2.$$

The calculations for the data along with the expected frequencies for raw data from Anderson (1958) have been given below.

Table 45.1: EDF values for the Anderson Data

Head length	*Expected x F (y)*	*EDF Values*	*Head Length*	*Expected F(y)*	*EDF Values*
163	0.00135	0	188	0.5	0.458333
174	0.00298	0	188	0.598706	0.5
174	0.00621	0	189	0.691462	0.666667
175	0.012224	0.041667	190	0.773373	0.791667
176	0.02275	0.041667	191	0.841345	0.875

(contd...)

Head length	*Expected x F (y)*	*EDF Values*	*Head Length*	*Expected F(y)*	*EDF Values*
176	0.040059	0.041667	192	0.89435	0.958333
179	0.066807	0.041667	192	0.933193	0.958333
181	0.10565	0.041667	195	0.959941	0.958333
181	0.158655	0.166667	195	0.97725	0.958333
183	0.226627	0.166667	197	0.987776	0.958333
183	0.308538	0.291667	197	0 99379	1
186	0.401294	0.458333	208	0.99702	1

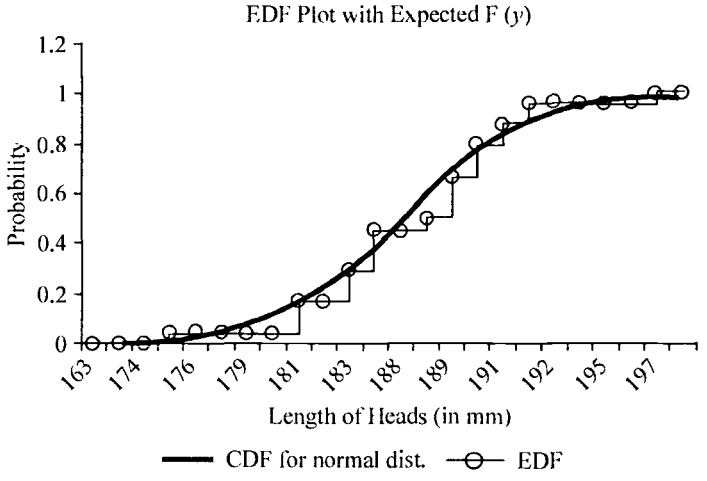


Fig. 45.1. *EDF plot for the Anderson Data*

From the plot it appears that the data is a considerably good fit to the normal distribution.

45.3. Axes

X Axis: Along this axes we consider the values of the observations. In this case we consider the length of heads in mm.

Y Axis: Along this axis we consider the values of the cumulative probabilities and the corresponding values of the empirical distribution function. Thus the axis ranges from 0 to 1.

45.4. Advantages

(*a*) EDF plot provides a quick check to normality of the data, or of any other distribution that the data may follow.

(*b*) It easy to draw and simple to understand.

45.5. Disadvantages

(*a*) The problem with this type of plot is that the conclusions regarding the goodness of fit test taken by different users may be subjective as the proximity of the EDF curve and the CDF curve may be differently interpreted by the users. A suggestion can be to develop a proper confidence bands around the CDF curve and so the step function lying within the confidence band may act as the indicator for a good fit.

(*b*) It is better if the theoretical distribution is fully specified along with the parameters.

(*c*) The plot remains scale sensitive *i.e.*, by changing the scale along the Y-axis one can manipulate the conclusion related to the goodness of fit.

(*d*) The plot is less frequent in statistical packages.

45.6. Related Techniques

(*a*) Chi-square Test for the Goodness of Fit;

(*b*) Anderson-Darling Test;

(*c*) Wilks Shapiro Test;

(*d*) Normal Probability Plot; and

(*e*) Probability Plot.

46. EMPIRICAL DISTRIBUTION FUNCTION PLOT (WITH DOKSUM BOUNDS)

46.1. Definition and Description

The plot is similar to that of the EDF plot discussed earlier, only that this step goes forward to generate a confidence band around the CDF curve such that the disadvantage of the EDF plot can be overcome.

We know that the famous non– parametric test, Kolmogrov – Smirnov test, that is used to test if a given random sample comes from a specific distribution, is based on the EDF and the test statistic is given by

$$D = \text{Sup} \mid F_n(y) - F(y) \mid$$

If the CDF is plotted, then a confidence band can easily be drawn around it based on the $K - S$ statistic. Kotz and Johnson (1988) give a hint about the development of such bands. But on plotting the band it is seen that such bands are unnecessarily broad in the tails. Keeping this issue in mind Doksum (1977) improved the K–S Statistic by dividing it by a factor $\{F(x)\,(1 - F(x))\}^{1/2}$. This factor acts as variance equalizer and the band thus generated would be slightly wider in the middle and much narrower at the tails compared to the first band. Accordingly the following expression may be derived.

$$1 - \alpha = P\left[\text{Sup}\frac{|F_n(y) - F(y)|}{[F(y)(1-F(y))]^{1/2}} \le D_\alpha(n)\right]$$

$$= P\left[\frac{|F_n(y) - F(y)|}{[F(y)(1-F(y))]^{1/2}} \le D_\alpha(n) \,\forall\, y\right]$$

$$= P\left[-D_\alpha^{(n)} \le \frac{F_n(y) - F(y)}{[F(y)(1-F(y))]^{1/2}} \le D_\alpha(n) \,\forall\, y\right]$$

$$= P\left[F(y) - \sqrt{F(y)(1-F(y)}\, D_\alpha(n) \le F_n(y)\right.$$

$$\left. \le F(y) + \sqrt{F(y)(1-F(y)}\; D_\alpha(n) \;\forall\, y\right]$$

Thus, $[F(y) - \sqrt{F(y)(1-F(y))}\, D_\alpha(n),\ F(y) + \sqrt{F(y)(1-F(y))}\;\; D_\alpha(n)]$ is the $100\,(1-\alpha)\%$ Doksum confidence band for $F_n(y)$.

Thus for different values of the variable one may plot the EDF, CDF of the theoretical distribution and the Doksum bound deduced for a given level of significance. If the EDF lies within the band then the data may be considered as a good fit to the theoretical distribution.

46.2. Working Data and Calculations

The data collected for the purpose is taken from Anderson (1958) which pertains to the head length of a group of individuals. To check if the observations can be considered to have come from normal distribution we draw the EDF plot and compare it with the corresponding CDF. It is better if the distribution is completely specified otherwise the parameters are estimated using their corresponding maximum likelihood estimators *i.e.*,

$$\hat{\mu} = \bar{x} = \frac{1}{n}\sum x_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum (x_i - \bar{x})^2.$$

The calculations for the data along with the EDF values and the Doksum bounds for raw data from Anderson (1958) have been given below.

Table 46.1: EDF and Doksum bounds of the data

Head length	*Expected F(y)*	*EDF*	*Lower Doksum Bounds*	*Upper Doksum Bounds*
163	0.00135	0	–0.00856	0.011264
174	0.00298	0	–0.01174	0.017697
174	0.00621	0	–0.015	0.02742
175	0.012224	0.041667	–0.01744	0.041894
176	0.02275	0.041667	–0.01751	0.063009
176	0.040059	0.041667	–0.01289	0.093006
179	0.066807	0.041667	–0.00061	0.134223
181	0.10565	0.041667	0.022655	0.188645
181	0.158655	0.166667	0.06001	0.257301
183	0.226627	0.166667	0.113592	0.339663
183	0.308538	0.291667	0.183827	0.433248
186	0.401294	0.458333	0.26895	0.533637
188	0.5	0.458333	0.365	0.635
188	0.598706	0.5	0.466363	0.73105
189	0.691462	0.666667	0.566752	0.816173
190	0.773373	0.791667	0.660337	0.886408
191	0.841345	0.875	0.742699	0.93999
192	0.89435	0.958333	0.811355	0.977345
192	0.933193	0.958333	0.865777	1.000608
195	0.959941	0.958333	0.906994	1.012887
195	0.97725	0.958333	0.936991	1.017508
197	0.987776	0.958333	0.958106	1.017445
197	0.99379	1	0.97258	1.015001
208	0.99702	1	0.982303	1.011737

46.3. Axes

X Axis: Along this axes we consider the values of the observations. In this case we consider the length of heads in mm.

Y Axis: Along this axis we consider the values of the cumulative probabilities, the corresponding values of the empirical distribution function, values of the Doksum bounds.

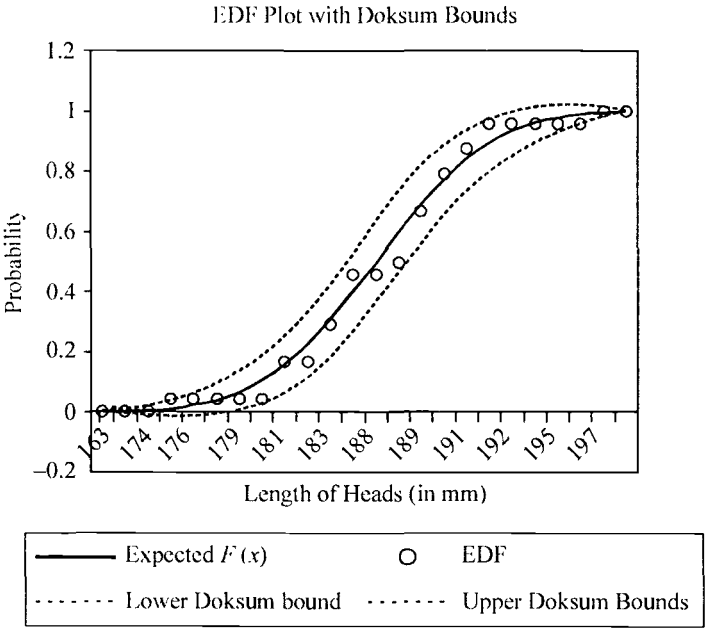


Fig. 46.1. *An EDF plot along with the Doksum Bounds*

From the plot it appears that the data is a considerably good fit to the normal distribution.

46.3. Advantages

(*a*) The plot provides a quick check to normality of the data, or of any other distribution that the data may follow at any required level of significance.

(*b*) It is easy to draw and simple to understand.

(*c*) The use of bands overcomes the disadvantage related to the subjective approach of different users in the interpretation of the plot.

(*d*) The plot does not remain scale sensitive.

46.4. Disadvantage

(*a*) It is better if the theoretical distribution is fully specified along with the parameters as the critical region of the Kolmogrov-Smirnov test may not remain valid if the parameters are estimated from the data.

(*b*) Probably none of the statistical packages provide an option to draw the plot.

46.5. Related Techniques

(*a*) Chi-square test for the goodness of fit;
(*b*) Anderson-Darling test;
(*c*) Wilks Shapiro test;
(*d*) Normal Probability Plot; and
(*e*) Probability Plot.

47. ERROR BAR PLOT

47.1. Definition and Description

An error bar plot is a graphical data analysis technique for showing the error in the dependent variable and, optionally, the independent variable in a standard x-y plot. As in a standard x-y plot, the vertical axis contains a dependent variable while the horizontal axis contains an independent variable. In addition, it contains error bars in the vertical direction and/or if desired along the horizontal direction for visualizing the error due to the independent variable. The error bars can be either symmetric or asymmetric about the point. The error bars are used to indicate the estimated error in a measurement. Errors bars indicate the uncertainty in the x and/or y values. The most common error bar plot is one in which the errors are in the y-values. For each value of the independent variable a number of replications of the dependent variable are taken the means and standard errors of the dependent variables (y) for each value of the independent variable (x) are computed. Then $(x_i, \bar{y}_i)$ are plotted and around each point the bars are allowed to extend from $(x_i, \bar{y}_i + \text{SE}(\bar{y}_i))$ to $(x_i, \bar{y}_i + \text{SE}(\bar{y}_i))$.

47.2. Working Data

For drawing the jittered plot the following data is considered originally from Snedecor (1956) which gives the birth weights (in lbs.) of Poland China pigs for four litters in pounds.

Table 47.1: Birth weight of Poland China Pigs in Pounds

Liter 1	*Liter 2*	*Liter 3*	*Liter 4*
2.0	3.5	3.3	3.2
2.8	2.8	3.6	3.3
3.3	3.2	2.6	3.2
3.2	3.5	3.1	2.9
4.4	2.3	3.2	2.0
3.6	2.4	3.3	2.0
1.9	2.0	2.9	2.1
3.3	1.6	3.4	—
2.8	—	3.2	—
1.1	—	3.2	—

47.3. Axes

X Axis: It can represent the values of the independent variable. For this data set we consider the Liters along the x-axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took Birth weights along the y axis.

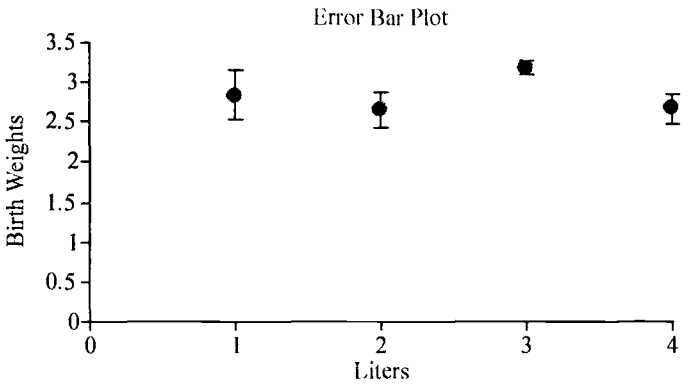


Fig. 47.1. *An Error Bar plot for the data in Table 47.1*

47.4. Advantages

(*a*) The plot is used for comparison of the central values of a number of groups.

(*b*) The plot can also be used to view the spread of the data around the central value. Hence, the plot can be used as a tool for comparison of central values and deviation about the central value.

(*c*) The plot can be drawn without the use of color.

(*d*) This is commonly available in most statistical software including MS Excel.

47.5. Disadvantages

(*a*) The plot can only visualize the difference but it is difficult to infer from it, especially in case of not so pronounced difference.

(*b*) If the mean ± SE of the response variable for a particular group is very large compared to other groups then the comparison of the dispersion across the groups become difficult to perform.

47.6. Related Techniques

(*a*) Jittered Plot;

(*b*) ANOVA;

(*c*) *t* test; and

(*d*) F test.

48. EWMA CONTROL CHART

48.1. Definition and Description

The full form of ERMA is exponentially weighted moving average. Some authors also call it as geometric moving average control charts. The chart is used to detect if the measurement process has gone out of statistical control. It is more efficient than any other control chart in detecting small shifts in process mean.

Here k samples of the manufactured product are collected of same size 'n' (say) and the measurements are taken. Let x_{ij} be the value of the i^{th} ($i = 1, 2, \ldots, n$) observation in the j^{th} ($j = 1, 2, \ldots, k$) sample. Next, we compute the mean ($\bar{x}_i$) for each of the samples.

In the chart we plot y_i against i.

Where, $$y_i = p\bar{x}_i + (1-p)\, y_{i-1} \qquad \ldots(48.1)$$

with, p = a fraction between 0 and 1.

$\bar{x}_i$ = subgroup average at time t.

y_0 = overall mean.

Thus the value of y_i is the sum of two weighted averages. From the expression it is clear that as the value of p tends to 1 the effect of prior mean (y_{i-1}) diminishes and vice versa. In case $p = 1$, then the EWMA Control Chart becomes equivalent to $\bar{x}_i$ - chart.

Here we have,

$$\text{Upper Control Limit} = UCL = \bar{\bar{x}} + 3.692\hat{\sigma}\sqrt{\frac{p}{k(2-p)}}$$

$$\text{Lower Control Limit} = LCL = \bar{\bar{x}} - 3.692\hat{\sigma}\sqrt{\frac{p}{k(2-p)}}$$

48.2. Working Data

The data for this purpose is the life in hours of cells after complete charging obtained from a local manufacturer of cells. The data consists of 15 samples each of size 5.

Table 48.1: Life in hours of cells after full charging

Sample No	*Life of Cells in Hours*					*Mean* ($\bar{x}_i$)
1	30.5	34.6	20.2	29.0	30.7	29
2	23.9	22.1	38.7	39.8	27.5	30.4
3	31.1	23.5	31.4	27.3	39.6	30.58
4	24.3	39.4	30.9	21.9	25.2	28.34
5	20.7	20.6	37.7	26.3	32.3	27.52
6	36.6	27.1	20.1	33.8	29.9	29.5

(contd...)

Sample No	*Life of Cells in Hours*					*Mean* ($\bar{x}_i$)
7	37.7	27.9	20.2	25.7	36.2	29.54
8	26.7	38 4	22 4	29.4	24.1	28.2
9	29.7	20.3	30.6	35.2	36.1	30.38
10	22.8	33.3	32 3	33.0	27.2	29.72
11	25.9	22.4	29.3	28.2	21.6	25.48
12	30.1	31.6	24.4	27.1	27.7	7.24

Thus, we have,

$$\bar{\bar{x}} = \frac{\sum \bar{x}_i}{k} = 28.969$$

$$\hat{\sigma} = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(\bar{x}_i - \bar{\bar{x}})^2} = 1.5187$$

Taking, $p = 0.7$ we have,

Upper Control Limit = $UCL = \bar{\bar{x}} + 3.692\hat{\sigma}\sqrt{\dfrac{0.7}{12(2-0.7)}} = 30.2$

Lower Control Limit = $LCL = \bar{\bar{x}} - 3.692\hat{\sigma}\sqrt{\dfrac{0.7}{12(2-0.7)}} = 27.7$

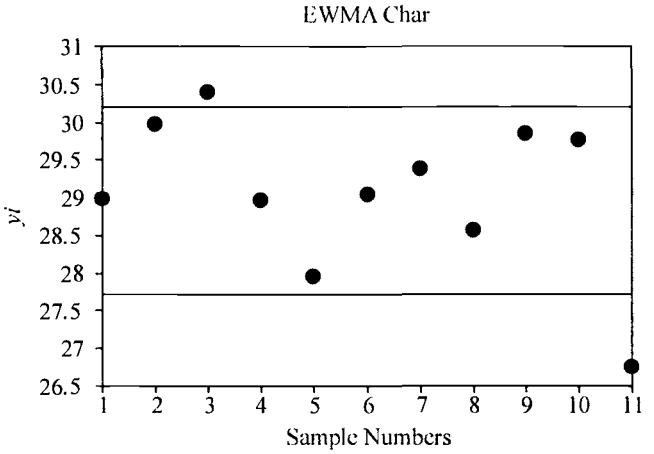


Fig. 48.1. *EWMA Plot for the data provided in Table 48 1*

Thus, we see that in the plot two points are outside the control limit implying that the system is out of control.

48.3. Axes

X Axis: The axis is used to represent the sample numbers.

Y Axis: This axis is used to represent the values y_i obtained from equation (48.1).

48.4. Advantages

(*a*) The plot is simple to draw and easy to interpret.

(*b*) The chart is used to discover the existence of any assignable cause of variation.

(*c*) The chart is successful even in case of small samples.

48.5. Disadvantages

(*a*) The calculations involved are not very simple.

(*b*) The choice of p in (48.1) is subjective. Different users may use different values of p and accordingly reach to different conclusions for the same set of data.

48.6. Related Techniques

(*a*) $\bar{x}$ Chart;

(*b*) Cusum Chart; and

(*c*) R Chart.

49. EXTREME PLOT

49.1. Definition and Description

The extreme plot is used to detect if the largest value of different sub-samples differs from each other. This can also be used to study variation in the highest order statistics across the different sub-samples. The plot is a scatter plot between the sub-sample identification number and the largest observation of the sub-sample. In case of large variation in the extreme values the points will be scattered, otherwise they will form an approximate straight line parallel to X axis.

49.2. Working Data

The data used for drawing the plot comprises of 10 sub-samples of random numbers drawn from the uniform distribution U (0, 1). Each sub-sample comprises of 7 observations.

Table 49.1: 10 sub-samples of size 7 each drawn from a U [0, 1] population

Sub sample No.	1	2	3	4	5	6	7	8	9	10
	0.42	0.03	0.81	0.81	0.49	0.58	0.44	0.80	0.86	0.77
	0.77	0.95	0.07	0.86	0.87	0.29	0.30	0.07	0.95	0.65
	0.34	0.98	0 45	0.46	0 55	0.44	0.59	0.18	0.90	0.28
	0.12	0.10	0.97	0.22	0.96	0.02	0.82	0.16	0.56	0.56
	0.44	0.48	0.32	0.99	0.64	0.25	0.31	0.22	0 45	0.20
	0.20	0.39	0.38	0.58	0.63	0 57	0.79	0.27	0.69	0.44
	0.61	0.32	0.44	0.29	0.56	0.32	0.71	0.02	0.87	0.88
Max.	0.77	0.98	0.97	0.99	0.96	0.58	0.82	0.80	0.95	0.88

49.3. Axes

X Axis: The axis is used to represent the sub-sample identification number.

Y Axis: This axis is used to represent the extreme value of the sub-sample.

49.4. Uses

(*a*) The plot provides a means to compare the extreme values for various subsamples.

(*b*) It helps in to understand the distributional pattern of the n^{th} order statistics.

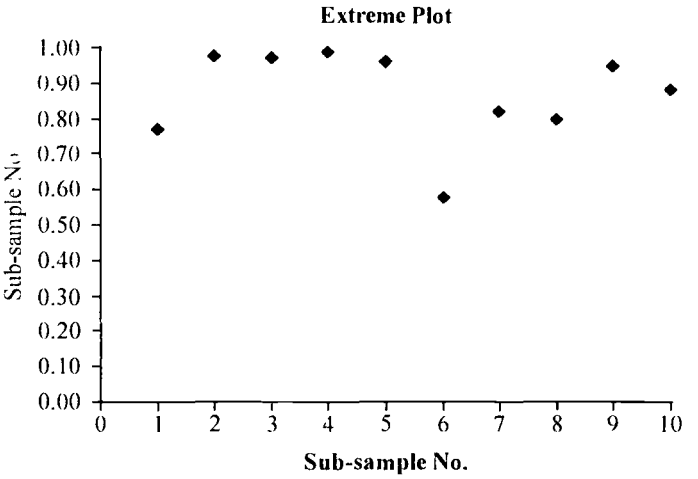


Fig. 49.1. *An Extreme plot for the data provided in Table 49.1*

Here we see that except sub-sample number 6, the maximum values of other sub-samples does not show much difference.

49.5. Related Techniques

(*a*) Mean Plot;

(*b*) Standard Deviation Plot; and

(*c*) Linear Correlation Plot.

50. FISH BONE CHART

50.1. Definition and Description

This technique is a semi-graphical tool used for analyzing complex problems which seem to have numerous and interrelated causes. It explores the relationship between the problem and its causes (by category) and presents them visually. It may also be called an Ishikawa Diagram, named after the Japanese Professor Kaoru Ishikawa, who first used this technique in 1943.

At the initial stage define the goal. Write the goal on the far right side of your page and draw a long horizontal arrow pointing towards it. This line may be considered as the back-bone of the plot and leads to the goal. From the horizontal arrow, the user may be able to branch off some major and minor steps or problems that may be faced to reach the goal. First one must identify the most significant causes, or potential causes for reaching the goal. These will form the main branches form the "backbone arrow".

For detailed analysis, and drawing more sub-problems one may draw sub-branches from the branches and point out more problems.

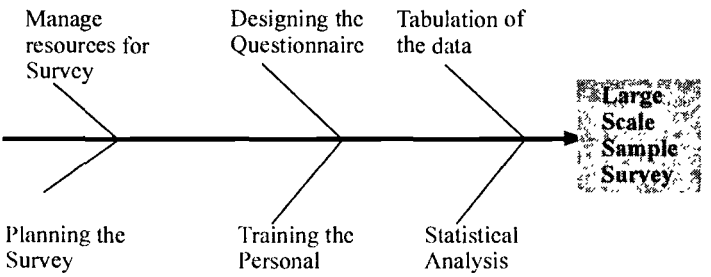


Fig. 50.1. *A Fishbone chart showing the problems related to a sample survey*

50.2. Uses

(*a*) It diagram can be used to understand the present position of a project and what is the next immediate target.

(*b*) It can be adapted for structuring a brainstorming session.

(*c*) It different surveys the technique can be used in explaining the enumerators and other officials associated with the survey about potential or actual problems and in overcoming the factors responsible for these causes.

50.3. Related Techniques

(*a*) Tree Diagram; and

(*b*) Flow Chart.

51. FOUR FOLD DISPLAY

51.1. Definition and Description

A four fold display is used to display the cell frequency in a 2 × 2 contingency table. This was introduced by Friendly (1994a, 1994b). The display is done in such a way that it also displays the odds ratio of the contingency table. A 2 × 2 contingency table is generally in the following form:

		Category A	
		I	II
Categorgy	I	θ_{11}	θ_{12}
	II	θ_{21}	θ_{22}

The odds ratio is given by $\theta = \frac{\theta_{11}}{\theta_{12}} \times \frac{\theta_{22}}{\theta_{21}}$. ...(51.1)

If $\theta = 1$, then it implies that the categories A and B are mutually independent. In this display the frequency in each cell is shown by a quarter circle, whose radius is proportional to $\sqrt{O_{ij}}$, (where O_{ij} is the observed frequency corresponding to the i^{th} row and j^{th} column) so the area of each quarter is proportional to the observed frequency. The quarter circles are then colored or shaded depending on the direction and magnitude of the Pearsonian residual. If there is an independence between the categorical variables then the odds–ratio (θ) tends to one, and the diagonally opposite quarters of the four fold display are approximately of the same size as that of the quarters in the other diagonal. However in case of association between variables the odd ratio is far from 1 and in such a case the diagonally opposite cells in one direction differs in size from those of the opposite direction.

However, some of the four fold displays the raw frequencies are not used for plotting instead the frequencies are standardized using iterative proportional fitting. This is done in order to make the quarter circles comparable.

51.2. Working Data

The table given below is a 2 × 2 table, taken from Walker and Lanes (1991). The table shows death or hospitalization due to asthma cross classified by Fenoterol use or not.

Table 51.1: Asthma deaths, and Fenoterol use from Walker and Lanes (1991)

Category 1 / *Category 2*	*Fenoterol Used*	*Fenoterol Not Used*
Deaths	34	50
Hospitalized	151	213

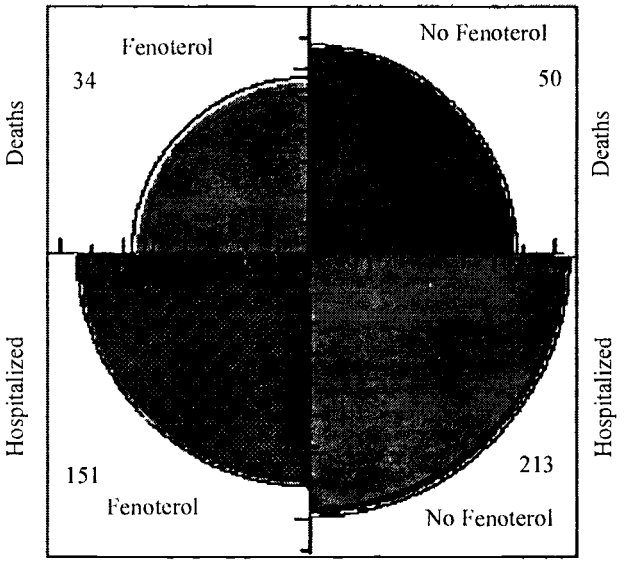


Fig. 51.1. *A four fold display for the raw frequencies given in Table 51.1*

From the display we can see that the diagonally opposite cells in one direction show some difference in size in the opposite direction. This indicates that there is some association between the two categories viz., death or hospitalization due to asthma and Fenoterol use.

51.3. Axes

For this plot no axes is required.

51.4. Advantages

(*a*) The plot is used for the visualization of data in case of data arranged in the form of contingency table.

(*b*) The plot can be used to understand the dependence if any between two different attributes.

(*c*) The plot is simple to draw and its interpretation is also easy.

51.5. Disadvantages

(*a*) The criterion of independence is sometimes difficult to be assessed visually and the user in all cases may not be confident in his decision.

(*b*) This display does not support any comparison between the observed and expected frequencies.

(*c*) The display is restricted to 2×2 contingency table only.

51.7. Related Techniques

(*a*) Chi-square test for independence in contingency table;
(*b*) Association Plot;
(*c*) Sieve Diagram; and
(*d*) Mosaic Display.

52. FREQUENCY POLYGON/CURVE

52.1. Definition and Description

Frequency polygon or frequency curve is a very commonly used graphical technique both for grouped and ungrouped frequency distribution. In the case of an ungrouped frequency distribution, we take the values of the variate along the horizontal axis and plot the corresponding frequencies along the vertical axis by using suitable scales. These points are then added together. If the points are added by a free hand smooth curve then it is called as frequency curve but if the points are added by straight line then it is called as frequency polygon. Frequency polygons are also used for representing grouped frequency distributions, but only when all the class intervals are equal. In such a case, frequency of each class is plotted against the mid-values of that class interval. The polygon should be brought down at both ends to the X axis by joining it to the class marks of the nearest empty class at each end of the distribution.

52.2. Working Data

The working data is a hypothetical data of a grouped frequency distribution, showing the wages of a group of employees.

Table 52.1: A frequency distribution of wages of a group of employees

Wages (in Rs.):	50–60	60–70	70–80	80–90	90–100	100–110
No. of employees.	8	10	16	14	10	5

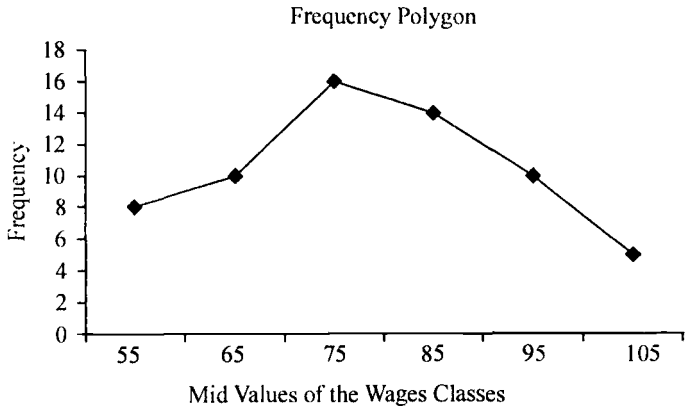


Fig. 52.1. *A frequency polygon representing Table 52.1*

52.3. Axes

X Axis: It can represent the mid values of the classes in case of grouped frequency distribution. For this data set Wages (in Rs.) is taken along the X-axis.

Y Axis: The vertical axis is used for representing the frequencies. For this case we took the number of employees along the Y-axis.

52.4. Advantages

(*a*) The frequency polygon gives us an approximate idea of the shape of the frequency curve.

(*b*) A rough idea about the dispersion of the variable may also be understood.

(*c*) From the diagram one may get an idea about the skewness and kurtosis of the distribution.

(*d*) Use of color for drawing such a plot is not essential.

52.5. Disadvantages

(*a*) This plot cannot be used for detection of any measure of central tendency.

(*b*) The plot can only be drawn in case the class intervals are of equal width.

(*c*) The plot cannot be drawn in case of open end classes.

52.6. Related Techniques

(*a*) Histogram;

(*b*) Cumulative Frequency Curve; and

(*c*) Frequency Distribution.

53. GLYPH PLOT

53.1. Definition and Description

The Glyph plot is a graphical technique that can be used for representing three variables or even more variables in two dimensions each of which are numerical variables. The word 'Glyph' refers to a relatively complex plotting symbol that can be made to vary in different ways in order to visualize several variables simultaneously. This plot can be considered as an extension of a scatter plot. Here we choose two primary variables that are represented along X-axis and Y-axis respectively. The (x, y) points are plotted in the graph paper but in the forms of small but filled squares (■) or any other symbol. Then from each of the plotted points tails are drawn the length of which are proportional to the value of the third variable. The tails are generally drawn diagonally downwards of the squares and the length of the tails are proportional to the value of the third variable. The ray or tail that comes out from the plotted point must have a minimum length. The minimum length of the glyph corresponds for the minimum value of the data.

53.2. Working Data

The data used for drawing the glyph plot given in Table 17.1. Here three of the variables viz, Height, Weight and Age is used in drawing the graph in Figure 17.1 that follows. Thus the plot is used for representing three variables each of which is a numerical variable.

53.3. Axes

X-Axis: It can represent the values of the variable that we suspect may have a relation with the response variable. Here heights of the individuals are taken along the axis.

Y-Axis: The vertical axis consists of that variable which we consider as the response variable. Here weights of the individuals are taken along the axis.

The third variable is represented by a whisker that comes out from the plotted point and is proportional to its numerical value.

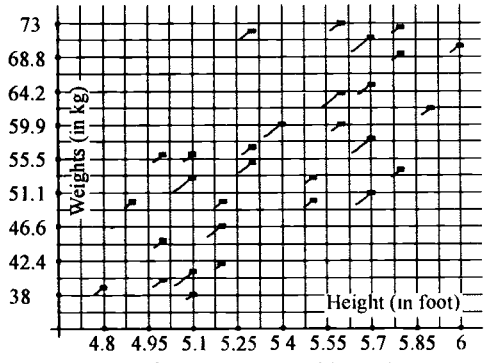


Fig. 53.1. *A glyph plot for the data in Table 17.1 representing Height, Weight and Age of 30 individuals*

53.4. Advantages

(*a*) The graphs can be used to study the relationship between X and Y variables.

(*b*) It can also be used to study the variation between X and Y in presence of a third variable.

(*c*) It is helpful in detection of outliers.

(*d*) It can be used for visualization of 3 numerical variables in two dimensions and can be extended for representation of more variables.

(*e*) This plot diminishes the chance of over-plotting as the whiskers are slim lines.

53.5. Disadvantages

(*a*) This graphical technique requires some calculations for converting the value of the third variable to proportional radius.

(*b*) The numerical variable that is represented by the whiskers is studied relative to each other. The values of such variables are not scaled to be read from the graph and thus the graph remains less informative about the third variable, 'Age' in this case.

53.6. Related Techniques

(*a*) Bubble Plot;

(*b*) Sunflower Plot; and

(*c*) Scatter Diagram.

54. GLYPH PLOT (CATEGORICAL)

53.1. Definition and Description

The categorical glyph plot is a graphical technique that can be used for representing of four variables three of which takes numerical values and the fourth one is a categorical variable. This plot can be considered as an extension of glyph plot. The glyph plot is drawn first in the manner discussed in the earlier section and then the plotting symbol *i.e.*, the glyphs are colored differently based on the value of the categorical variable. Here different colors are used to represent the different categories of the fourth variable, and accordingly we get glyph of different colors in the plot. Thus this plot can be used to represent four variables-three numerical variables and the fourth one is a categorical variable.

54.2. Working Data

The data used for drawing the glyph plots is given in Table 18.1. The data comprises of heights, weights, age and sex of 30 individuals collected non-randomly from some known individuals.

The glyph plot in Fig. 54.1 represents the first three variables of the Table viz. Height, Weight, Age and Sex.

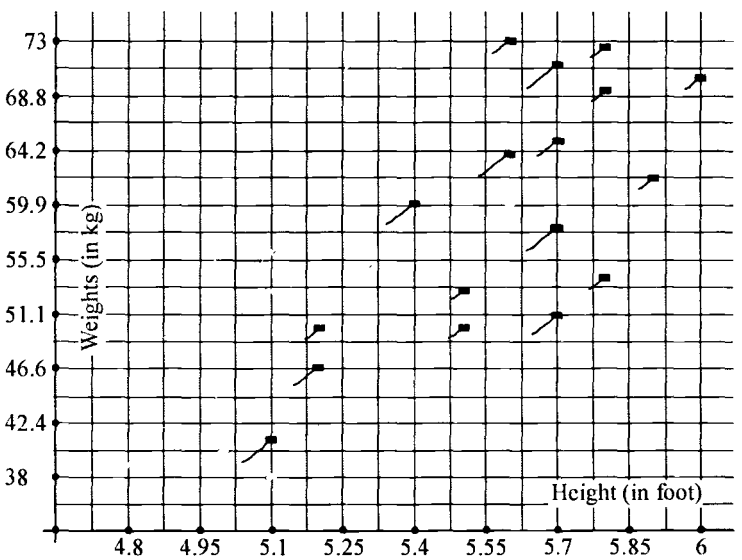


Fig. 54.1. *A categorical glyph plot for the data in Table 18.1*

54.3. Axes

X Axis: It can represent the values of the variable that we suspect may have a relation to the response variable. For this data set we consider weight to be related to the height of the individual and hence we have taken height of the individual along the X-axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. For this case we took weights along the Y-axis.

The third variable is age and is represented by the length of the whisker that comes out from the glyph. Thus the length of each whisker is proportional to the age of the subject.

The fourth variable is a categorical variable and is represented by the color of the glyph. Here a black glyph is used to represent the Males and the grey glyph is used to represent Females.

54.4. Advantages

(*a*) The graphs can be used to study the relationship between X and Y variables.

(*b*) It can also be used to study the variation between X and Y in presence of two other variables.

(*c*) It is helpful in the detection of outliers.

(*d*) It can be used for visualisation of four variables in two dimensions.

(*e*) It is used to understand how the value of the foreground variable varies across the various categories.

54.5. Disadvantages

(*a*) This graphical technique requires some calculations for converting the value of the third variable to proportional whisker.

(*b*) The numerical variable that is represented by the length of the whisker is studied relative to each other. The values of such variables are not scaled to be read from the graph and thus the graph remains less informative about the third variable, 'Age' in this case.

54.6. Related Techniques

(*a*) Categorical Bubble Plot;

(*b*) Categorical Sunflower Plot; and

(*c*) Categorical Scatter Plot.

55. HANGING ROOTOGRAM

55.1. Definition and Description

The hanging rootogram is a useful tool for checking the goodness of fit of a data to a particular distribution. Though a histogram can be used for the purpose of understanding the type of the distribution but it has some limitations. First, the histograms are drawn in such a manner that the diagram gets dominated by the larger frequencies compared to the smaller frequencies that lie in the tail areas. Secondly, from the overall position of the plots one cannot decide about the inherent form of the distribution. For example, if the histogram is bell shaped we cannot conclude that it is from normal distribution as the normal distribution is not the only distribution having such a form. In order to evaluate the distribution more exactly, one must introduce a comparison between the observed frequency and the theoretical frequency obtained from the assumed distribution in the plot. Thus, one may think of a histogram of observed frequencies plotted along with a curve showing the expected frequencies in such a way that the discrepancy can be visualized. But in such a case it becomes difficult to understand the difference between the two as (*i*) the smaller frequencies especially at the tails are difficult to visualize (*ii*) It becomes difficult to understand the difference between the histogram bars and the curve.

Keeping the two points in mind Tukey (1977) introduced the hanging rootogram. It is to root the frequencies and hang them to the theoretical curve and hence the name of the curve. In this plot, the following changes are made to a histogram:

(*a*) The frequencies are plotted on a square root scale, to make small frequencies relatively more prominent.

(*b*) The histogram bars start from the curve so that one may judge the differences more easily against a horizontal line.

55.2. Working Data

The data used for drawing the plot is taken from Jeffers (1978) which corresponds to the number of notices per day with its corresponding frequencies. The data is fitted to a Poisson distribution and the corresponding expected frequencies are computed.

Table 55.1: Data from Jeffers (1978) along with expected frequencies from Poisson distribution

Notices per day	0	1	2	3	4	5	6	7	8	9
Observed Frequency	162	267	271	185	111	61	27	8	3	1
Expected Frequency	127	273	295	212	114	49	18	6	2	0

55.3. Axes

X Axis: It represents the value of the independent variable. In the figure we take the number of notices per day is considered along the axis.

Y Axis: The vertical axis consists of the square root of the frequencies both observed and expected.

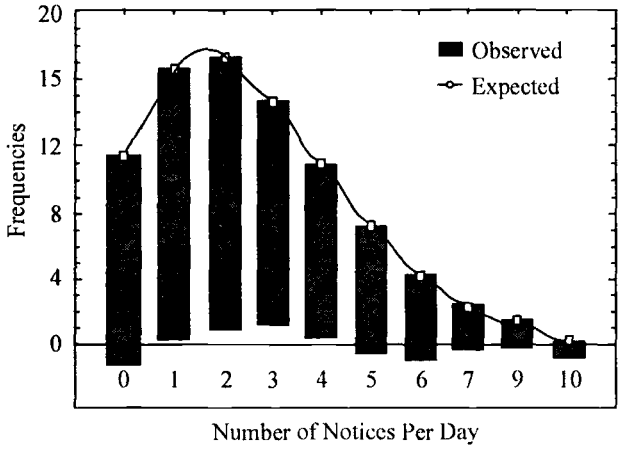


Fig. 55.1. *A hanging rootogram for the data in Jeffers (1978)*

55.4. Advantages

(*a*) The plot puts forward a visual impact on the quality of fit.

(*b*) The plot can be used to understand the pattern of variation between the observed and expected frequencies. For example, in the above figure if one looks at the base of the bars it appears that there is a periodic pattern noticed in the expected frequencies under the Poisson model. Ordinary goodness of fit tests does not reveal such facts.

55.5. Disadvantages

(*a*) Though the plot provides a visualization of the difference between the observed and expected frequencies but nothing can be inferred from it.

(*b*) It is difficult to compare two hanging rootograms to the same data but with expected frequencies from different distributions to understand a better fit.

55.6. Related Techniques

(*a*) Chi-square Goodness of Fit Test;

(*b*) Rootogram;

(*c*) Chigram; and

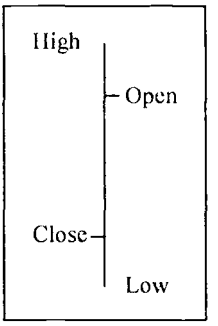
(*d*) Poissonness Plot.

56. HILO DIAGRAM

55.1. Definition and Description

The chart is used for the display of the highest and the lowest values of different data sets. Corresponding to each data set a vertical line is drawn parallel to the Y axis. The line starts from the minimum value and ends at the maximum value. Some small horizontal lines may be drawn from the vertical lines to signify the values of the mean, median etc. The different datasets are numbered 1, 2 etc. These integers are taken along the X axis and are used to identify the different datasets.

The plot can also be useful for share market experts to show the highest, lowest, closing as well as opening shares for several days in the same graph. For example:



56.2. Working Data

The data used for drawing the diagram is generated in MS Excel using the function for generating random numbers.

Table 56.1: Different data sets generated in Excel

Data Set No.	*Data Values*						*Highest*	*Lowest*	*Average*
1	9.08	8.81	8.87	9.29	8.81	5.46	9.29	5.46	8.39
2	9.51	5.84	8.12	4.43	0.56	2.52	9.51	0.56	5.16
3	5.29	9.24	0.18	3.61	9.93	3.60	9.93	0.18	5.31
4	1.05	0.60	0.37	6.18	4.01	0.78	6.18	0.37	2.16
5	0.42	1.40	0.94	4.44	0.31	8.96	8.96	0.31	2.74
6	4.80	5.58	9.95	3.07	5.50	3.68	9.95	3.07	5.43

56.3. Axes

X Axis: The axis is used to represent the data set identifies variable. Here time is considered.

Y Axis: This axis is used to represent the values of the response variable. Here we use the population along Y axis.

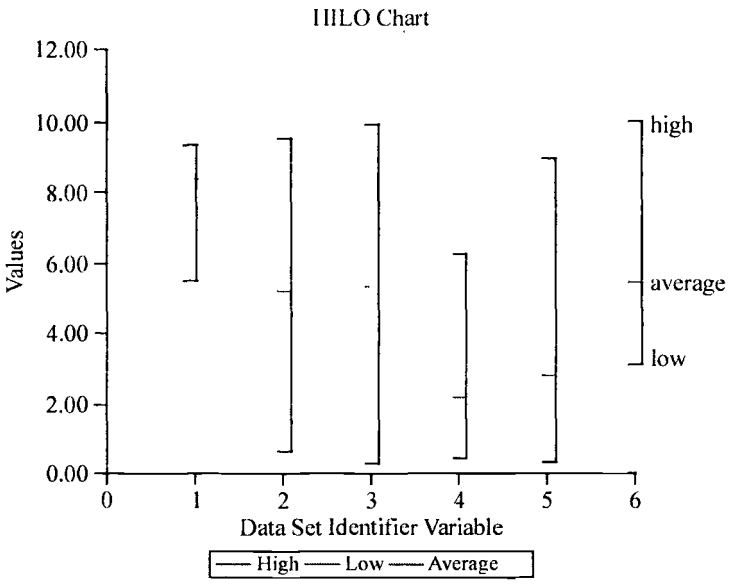


Fig. 56.1. *A HiLo diagram corresponding to the data in Table 56.1*

56.4. Uses

(*a*) The plot is simple to draw and the calculations are relatively easy.

(*b*) The chart is used to compare the means of several data sets.

(*c*) The chart is used for the purpose of comparison of the highest, lowest and the range of several data sets.

(*d*) The chart is useful in data representation of share markets.

56.5. Related Techniques

(*a*) Jittered Plot;

(*b*) Box Plot;

(*c*) Range; and

(*d*) Multiple Comparison.

57. HISTOGRAM

57.1. Definition and Description

One of the most commonly used methods of presenting frequency distribution of a continuous variable is known as histogram. In drawing a histogram, X axis is used for representing the class intervals and Y axis for frequencies. The histogram consists of a series of rectangles attached to one another. Here, one rectangle is used for each class. The heights of each rectangle are proportional to the frequency of the corresponding class. The width of the rectangles varies depending on the distance between the class boundaries and so they need not be equal.

57.2. Working Data

The working data is a hypothetical data of a grouped frequency distribution, showing the wages of a group of employees and is provided in Table 52.1.

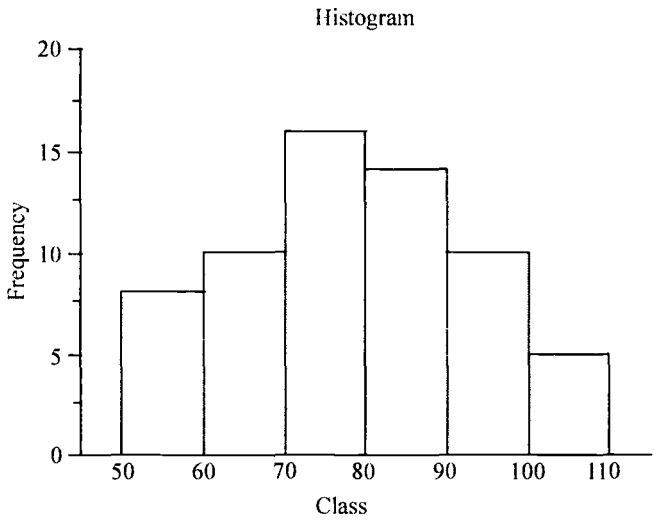


Fig. 57.1. *A histogram to the data in Table 52.1*

57.3. Axes

X Axis: It can represent the class intervals. For this data set Wages (in Rs.) is taken along the X-axis.

Y Axis: The vertical axis is used for representing the frequencies. For this case we took the number of employees along the Y-axis.

57.4. Advantages

(*a*) The series of rectangles in a histogram gives a visual representation of the relative sizes of frequencies in the various classes.

(*b*) From the diagram one may get an idea about the skewness and kurtosis of the distribution.

(*c*) Use of color for drawing such a plot is not essential.

(*d*) Useful in determining the mode of the frequency distribution.

(*e*) This plot can be used even if the class intervals are unequal.

57.5. Disadvantages

The plot cannot be drawn in case of open end classes.

57.6. Related Techniques

(*a*) Frequency Polygon;

(*b*) Cumulative Frequency Curve; and

(*c*) Frequency Distribution.

58. HISTORIGRAMS

58.1. Definition and Description

The historigram is same as that of line chart which is discussed in subsequent section of the book. The plot considers the value of the response variable along the Y-axis and that of the independent variable along the X-axis. The values of the dependent variable are plotted in the form of dots and then they are connected by free hand smooth curve. However, here the independent variable is always time variable. Also it may be noted that the time period should be long enough to be called as the historical period. Now 'historic period' is a relative term and is difficult to define. However for our convenience we may consider at least a decade as a historical period.

58.2. Working Data

The data used for the plot is the population of India for the various census years from 1901 to 2001 *i.e.*, the time period covers a century. The source of the data is Provisional Statistics, Census - 2001.

Table 58.1: Population of India for several census years

Census Year	*Population (Crores)*
1901	23.8
1911	25.2
1921	25.1
1931	27.8
1941	31.8
1951	36.1
1961	43.9
1971	54.8
1981	68.3
1991	84.3
2001	102.7

58.3. Axes

X Axis: The axis is used to represent the independent variable. Here time is considered.

Y Axis: This axis is used to represent the values of the response variable. Here we use the population along Y axis.

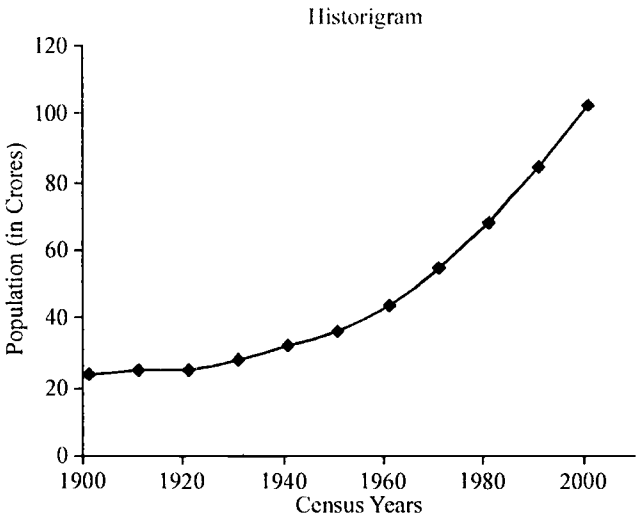


Fig. 58.1. *A Historigram for data in Table 58.1*

58.4. Uses

(*a*) The plot is simple to draw and no calculations are required.

(*b*) The chart is used to discover the existence of any trend in the data.

(*c*) The chart is useful in detecting the type of relationship of the response variable with the independent variable like, linear, quadratic, exponential etc.

58.5. Related Techniques

(*a*) Line Diagram;

(*b*) Run Sequence Plot;

(*c*) Scatter Diagram; and

(*d*) Moving Average Plot.

59. HOMOSCEDASTICITY PLOT

59.1. Definition and Description

In many statistical models the assumption of equality of variance is assumed the technical term of which is homoscedasticity. The homoscedasticity plot is used to check this assumption of constant variance across several subsets of a data. In the plot we consider the subset means across the X axis and the subset standard deviations across the Y axis. The points are plotted in the form of dots. The interpretation of the plot is also relatively simple. The more scattered are the points along the Y axis less valid is the assumption related to constant variance. However, if it is seen that with increase in the value of the mean the variance also increases then the data should be subjected to appropriate transformation before it is been plotted.

59.2. Working Data

The data used for drawing the diagram is generated in MS Excel using the function for generating random numbers.

Table 59.1: Different data sets generated in Excel

Data Set No.	*Data Values*						*Average*	*Standard Deviation*
1	9.08	8.39	8.87	9.29	8.81	5.46	8.39	1.45
2	9.51	5.16	8.12	4.43	0.56	2.52	5.16	3.37
3	5.29	5.31	0 18	3.61	9.93	3.60	5.31	3.71
4	1 05	2.16	0.37	6.18	4.01	0.78	2.16	2.38
5	0.42	2.74	0 94	4.44	0.31	8.96	2.74	3.40
6	4.80	5.43	9.95	3.07	5.50	3.68	5.43	2.43
7	5.03	5.25	4.74	5.17	4.85	4.17	4.87	0.39
8	4.35	4.65	4.05	4.48	4.19	3.95	4.28	0.27

59.3. Axes

X Axis: The axis is used to represent the sample means obtained from the different sub-sets.

Y Axis: This axis is used to represent the standard deviation of the subsets.

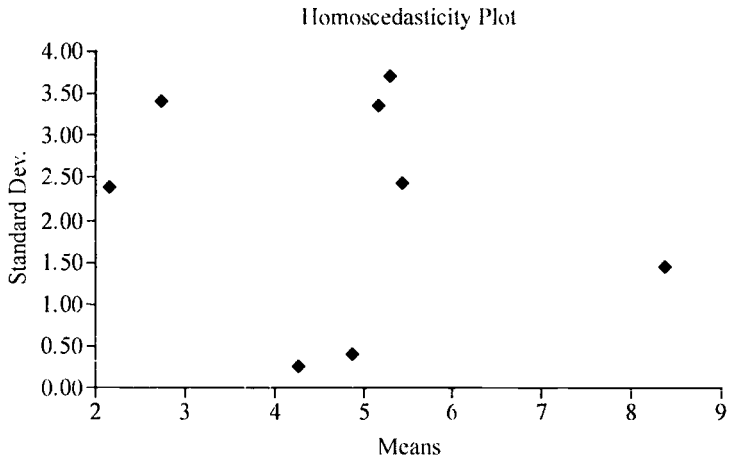


Fig. 59.1. *A homoscedasticity plot diagram corresponding to the data in Table 56.1*

The plot shows that the standard deviations are widely scattered across the vertical axis. So the assumption of constant variance cannot be assumed in this case.

59.4. Uses

(*a*) The plot is simple to draw and the calculations are relatively easy

(*b*) The chart is used to test the assumption of constant variance.

(*c*) The relation between mean and variance in the data sets can also be understood.

(*d*) The chart detects the need of any transformation in the data for stabilizing the variance.

59.5. Related Techniques

(*a*) Box Cox Homoscedasticity Plot;

(*b*) DOE Standard Deviation Plot;

(*c*) DOE Mean Plot; and

(*d*) Multiple comparison.

60. I PLOT

60.1. Definition and Description

The I plot is used to visualize the difference between the different levels of a 1 factor experiment. The values of the response variable are considered along the Y axis and a level identifier variable is considered along the X axis. In case each level is replicated a number of times one may use the mean of each level for plotting against the level identifier variable considered along the X axis. Some authors also suggests the calculation of the standard deviation for each level in case of a replicated experiment and accordingly *plot mean* ± *standard deviation* for each level. However, in such a case the plot produced takes the form of an Error Bar Plot discussed earlier.

60.2. Working Data

The data set for this purpose is taken from Rao (1948). The data refers to the weight of cork deposits in centimeters of 28 trees (here abridged to 10 trees only), in each of the four directions East, West, North and South.

Table 60.1: Cork deposits in centigrams in trees planted in different directions

	North	*East*	*South*	*West*
	72	66	76	77
	60	53	66	63
	56	57	64	55
	41	29	36	35
	32	32	35	36
	30	35	34	26
	39	39	31	27
	42	43	31	25
	37	40	31	25
	33	29	27	36
Mean	**44.2**	**42.3**	**43.1**	**40.5**

60.3. Axes

X Axis: In the axis one of the factors is considered which we may call the group identifier variable.

Y Axis: The vertical axis we consider the response variable. Here we consider the Cork deposits along the axis.

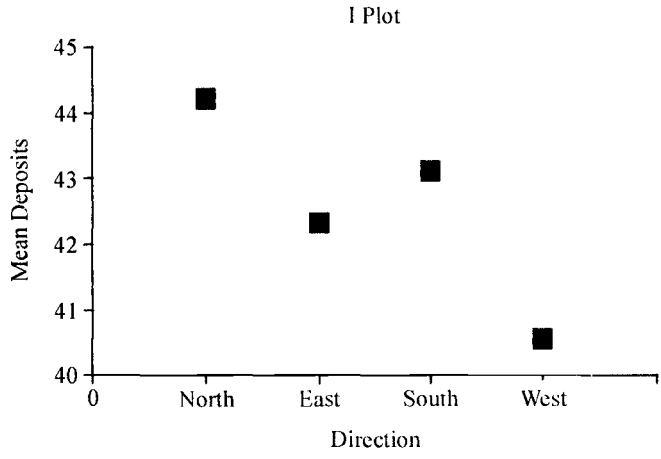


Fig. 60.1. *An I Plot for the data in Table 60.1*

From the figure we see that the maximum deposit of cork has taken place in the trees at the north and least has taken place in those in the west.

60.4. Uses

(*a*) To visualize the performance of different levels of a factor on the response variable.

(*b*) To identify the most dominating level.

(*c*) The plot can be drawn without the use of color.

(*d*) The plot is simple to draw and no calculations are required.

(*e*) The plot can accompany a one way classified data or an ANOVA for one way classified data.

60.5. Some Related Techniques

(*a*) Jittered Plot;

(*b*) Two-way Classified Data;

(*c*) Interaction Plot; and

(*d*) Block Plot.

61. INTERACTION PLOT

61.1. Definition and Description

The plot is used for the display of the effect of two factors in a response variable. In the plot the response variable is considered along the Y axis and one of the factors are considered along the X axis. The factor that is considered along X axis may be termed for our convenience as axis factor. The values of the response variable for a particular second factor are plotted in the graph and are connected by straight lines. Let this factor be termed as plotted factor. Next, the values of another response variable for another second factor are plotted in the graph and are again connected by another set of straight lines, but for a different color. The plot gives us an idea about how the response variable is affected by the interaction between two factors under consideration.

61.2. Working Data

The data set for this purpose is taken from Agresti (1989). The data shows the frequency of mental impairment in school going children cross classified by socio economic status of parents which increases from 1 to 6.

Table 61.1: Socio Economic status of parents and mental status of children

		Socio Economic Status of Parents					
		1	*2*	*3*	*4*	*5*	*6*
Mental Status	Well	64	57	57	72	36	21
	Mild Symptoms	94	94	105	141	97	71
	Moderate Symptoms	58	54	65	77	54	54
	Impaired Functioning	46	40	60	94	78	71

61.3. Axes

X Axis: In the axis one of the factors is considered which we may call as the axis factor. Here the socio economic status of parents is considered along the axis.

Y Axis: The vertical axis we consider the response variable. Here we consider the number of children along the axis.

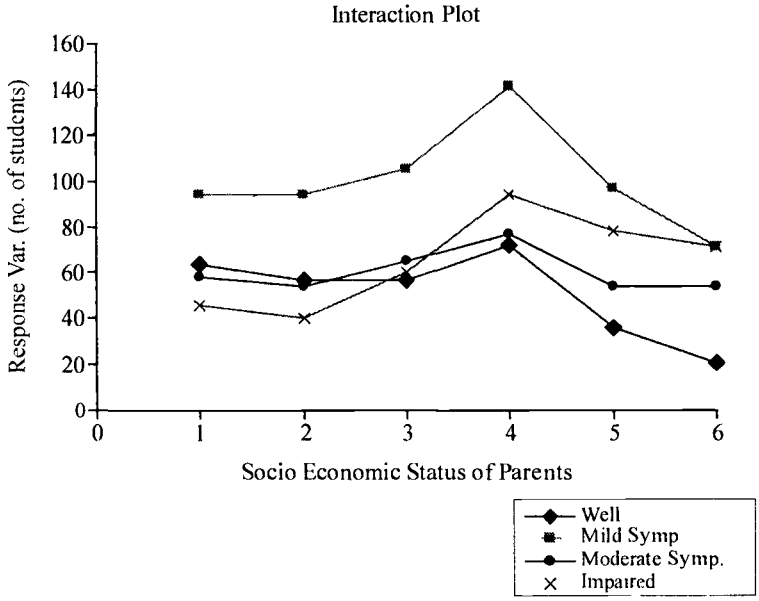


Fig. 61.1. *An Interaction Plot for the data in Table 61.1*

From the figure we see that most of the students have their parents at the socio-economic status '4'. The frequency of 'Impaired Functioning' are less amongst the students with parents from low socio-economic background and increases with the increase in the status of their parents. The 'Well' students have almost a decreasing trend with increasing status of parents.

61.4. Advantages

(*a*) To visualize the interaction of two factors in a response variable.

(*b*) To identify the most dominating class for each of the factors.

(*c*) The plot can be drawn without the use of color.

(*d*) The plot is simple to draw and no calculations are required.

61.5. Disadvantages

(*a*) The plot cannot be used to compare the means of different categories under different factors.

(*b*) The basic purpose of ANOVA is beyond the scope of the plot.

(*c*) The plot is simple but it is not always easy to interpret.

61.6. Some Related Techniques

(*a*) Jittered Plot;

(*b*) Two-way Classified Data; and

(*c*) Designs of Experiments.

62. JITTERED PLOT

62.1. Definition and Description

A jittered plot is commonly used to represent multiple number of groups in a graph. The groups are taken along the X axis which are been separated by a distance. The groups may be named or some identifying number can be used. Above each of the group identifier variables (along Y axis) the values of the response variable are plotted in the form of small circles. The circles are kept transparent in order to avoid overlapping. Another simple technique of avoiding overlapping is to generate a random number between [–0.2, 0.2] and add it to the group identifier variable before plotting. This disturbs the horizontal alignment of the points belonging to a particular group and hence acts a solution to the problem of overlapping in a group. The mean (or median) of the series is denoted by a small but thick straight line within the cluster of points in each group. This makes the means (or medians) readily visible and the means (or medians) across the groups can be compared.

62.2. Working Data

For drawing the jittered plot the data in Table 3.1 is considered.

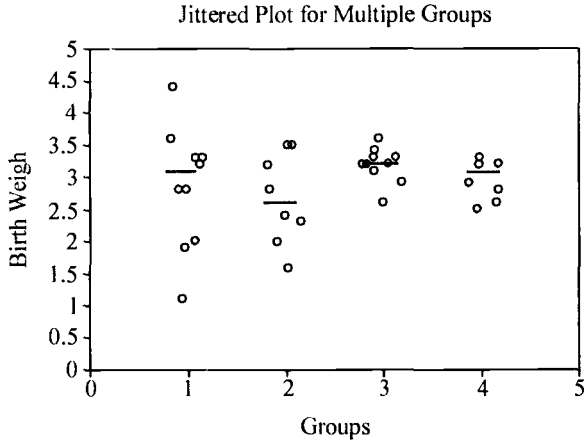


Fig. 62.1. *A jittered plot for data in Table 62.1*

From the jittered plot drawn above one can understand that the means of the second group is less, considered to that of the other groups. Also the spread of the data is seen to be maximum in the first group and minimum in the last one.

62.3. Axes

X Axis: It represents the different groups. The groups are generally coded using some integers called the group identifier variables.

Y Axis: The vertical axis consists of the values of the response variable. Here the Birth Weight are taken along the vertical axis.

62.4. Advantages

(*a*) The plot is used for comparison of the central values of a number of groups.

(*b*) The plot can also be used to view the spread of the data around the central value, range, symmetry around the central point and so on for each of the groups. Hence, the plot can be used as a tool for comparison of central values, deviation, range, concentration around central value and so on for multiple data sets.

(*c*) The plot can be drawn without the use of color.

62.5. Disadvantages

(*a*) The plot can only visualize the difference but it is difficult to infer from it, especially in case of not so pronounced difference.

(*b*) The plot is not commonly available in statistical packages.

(*c*) If the range of the response variable in a particular group is very large compared to other groups then the comparison of the dispersion across the groups become difficult to perform.

62.6. Related Techniques

(*a*) *t*-test;

(*b*) ANOVA Test;

(*c*) Block Plot; and

(*d*) F Test.

63. KAPLAN MEIER PLOT

63.1. Definition and Description

In many cases it is found that the outcome of interest is the time of occurrence of some events. In such cases it is not always possible to observe the events for a long duration of time. So the user has to censor the time and accordingly perform the analysis with censored data. Such a type of analysisis commonly encountered in medical science, engineering and other follow up studies.

Let us consider a set of random observation $X_1, X_2, \ldots, X_n$, representing failure time of n subjects, where $X_i \sim f_i(.)$, $i = 1, 2, \ldots, n$. But in the case mentioned above all the X_i's may not be observed completely because of loss due to follow up or because of time censoring.

Let C_i be the censoring time for the i^{th} subject. So we define,

$$T_i = min\,(X_i, C_i)$$
$$= I[T_i = X_i],$$

where I is an indicator function which takes the value 1 if $X_i < C_i$ and zero otherwise.

The Kaplan Meier estimator helps us to estimate the relaiability of a system for such censored data. Given n units, we obtain the values of $T_i = t_i$ for ($i = 1, 2, \ldots, n$) and they are ordered, $t_{(1)}, t_{(2)}, \ldots, t_{(n)}$, then the Kaplan Meier estimates are given by:

$$\hat{R}(t_i) = \prod_{j=1}^{i}\left(\frac{n-j}{n-j+1}\right)^{\delta_i}$$

The Kaplan Meier plot is the plot of $\hat{R}(t_i)$ versus the failure time.

63.2. Working Data

Ten components were studied. The study was stopped after 40 hours. By that time 7 components failed. The time of failure of the 7 components in hours were provided below: 10, 9, 34, 35, 6, 13 and 39. The corresponding Kaplan Meier plot is given in the Figure 63.1 below.

63.3. Axes

X Axis: The axis is used to represent the failure time of the components.

Y Axis: This axis is used to represent the values of the Kaplan Meier estimators *i.e.*, $\hat{R}(t_i)$.

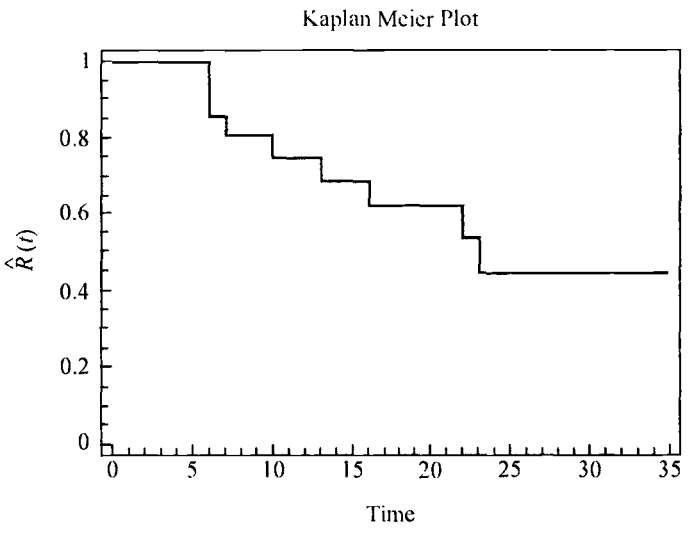


Fig. 63.1. *The Kaplan Meier Plot for data provided above.*

63.4. Uses

(*a*) The plot provides a means to compute the reliability of a system subjected to time censoring.

(*b*) It helps in the estimation of the CDF function of the failure time for time censored data.

63.5. Related Techniques

(*a*) Kaplan Meier Estimator;

(*b*) Time Censoring; and

(*c*) Duane Plot.

For further reading related to the plot the following references may be consulted:

- Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data.* New York: John Wiley & Sons.
- Lee, Elisa T. 1992. *Statistical Methods for Survival Data Analysis.* 2nd ed. New York: John Wiley & Sons.
- Cox, D. R. and Oakes, D. 1984. *Analysis of Survival Data.* London: Chapman and Hall.

64. LAG PLOT

64.1. Definition and Description

A lag plot is more generally used in time series compared to other sub-field of Statistics. It is used to check whether a data set or time series is random or not. Random data should not exhibit any identifiable structure in the lag plot. Non-random structure i.e. the existence of any trend would be understood by looking at the plot itself. Let Y_i ($i = 1, 2, ..., N$) be a time series data. A lag plot of lag h is a scatter plot between Y_i versus Y_{i-h}. In case the values of the time series is dependent on the lag values, then the scatter diagram so plotted will show a clear pattern, otherwise one may conclude that Y_i and Y_{i-h} are independent of each other. In case the plot has lag 1 then the scatter plot will be between Y_i versus Y_{i-1}.

64.2. Working Data

The data relates to deposits (in crores of rupees) of all the Regional Rural Banks (RRB) of India taken together from 1991 - 2004.

Table 64.1: Amount of Deposits in crores of Rupees in RRBs of India

Year	*DEPOSIT in crores*
1991	4989.24
1992	5867.83
1993	6938.14
1994	8826.51
1995	11150.01
1996	14187.9
1997	18032.01
1998	22189
1999	27065.74
2000	30051
2001	37027
2002	43200
2003	50098.34
2004	56350.08

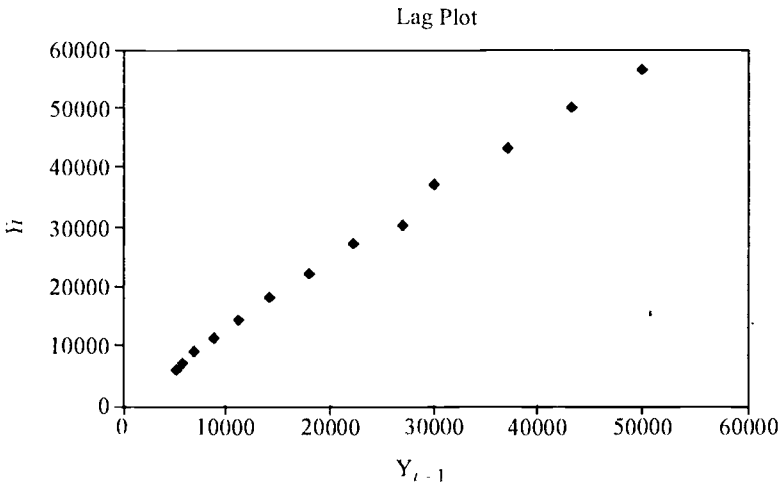


Fig. 64.1. *Lag Plot (of lag 1) for data in Table 64.1*

From the lag plot it is clear that there is a direct relation between deposits of a particular year with that of the previous year in the RRBs of India.

64.3. Axes

X Axis: The values of the lag variable are taken along the axis. In this case since the lag is 1 so we have taken the values Y_{i-1}.

Y Axis: The values of the time series are taken along the axes *i.e.*, the values of Y_i.

64.4. Advantages

(*a*) The plot is simple to draw and easy to interpret.

(*b*) It can be used to check the randomness of the data.

(*c*) To study the serial correlation in the data.

(*d*) Helps in deciding about a suitable model for a given time series data.

64.5. Disadvantage

Randomness is often a difficult task to perform by simple visualization. When the trend is very clear it is possible but this may not be the case always.

64.6. Related Techniques

1. Run Sequence Plot;
2. Scatter Diagram;
3. Serial Correlation; and
4. Auto Correlation.

65. LINE DIAGRAM

65.1. Definition and Description

The line diagram is generally used in Statistics to represent time series data. The line diagram helps us to understand how successive values of the variable are related to specified point of time. For constructing a line diagram, two axes of co-ordinates are taken perpendicular to each other, the horizontal one for time and the vertical one is used to represent the variable under consideration. The scale for each axis is then selected and the data are plotted. The plotting of variable values has been done against the different points of time. The successive points are now joined by straight line segments and the chart so obtained is called a *line diagram* for the given data.

65.2. Working Data

The data used for drawing the diagram gives the population of Assam (India) in thousands in the various census years.

Table 65.1: Population of Assam in different census

Year	1921	1931	1941	1951	1961	1971	1981	1991	2001
Poplation ('000)	4637	5560	6695	8029	10837	14625	18041	22414	26638

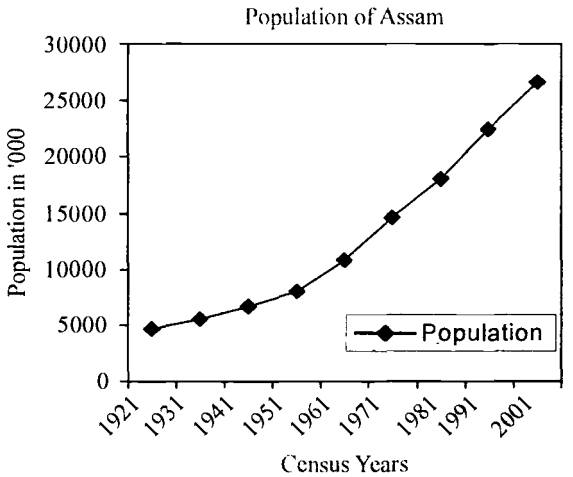


Fig. 65.1. *Line diagram corresponding to the data in Table 65.1*

65.3. Axes

X Axis: The axis is used to represent the independent variable. Here time is considered.

Y Axis: This axis is used to represent the values of the response variable. Here we use the population along *Y* axis.

65.4. Uses

(*a*) The plot is simple to draw and no calculations are required.

(*b*) The chart is used to discover the existence of any trend in the data.

(*c*) The chart is useful in detecting the type of relationship of the response variable with the independent variable like, linear, quadratic, exponential etc.

65.5. Related Techniques

(a) Run Sequence Plot;

(b) Scatter Diagram; and

(c) Moving Average Plot.

66. LINEAR INTERCEPT PLOT

66.1. Definition and Description

The linear intercept plot is used to detect if the intercept obtained from the least square linear fit between two variables differs across different sub samples. This can also be used to study variation in the intercept between two variables across the different sub-samples in the different categories. The plot is a scatter plot between the sub-sample identification number and the sub-sample intercept obtained from the linear fit. In case of large variation in intercepts the points will be scattered, otherwise they will form an approximate straight line parallel to X axis.

66.2. Working Data

The working data is hypothetical and comprises of the maximum temperature (in degree centigrade) and relative humidity (in %) collected for a period of fifteen days from five cities. The least square fit with maximum temperature as the independent variable and relative humidity as the dependent variable are shown as well for all the cities.

Table 66.1: Maximum temperature and relative humidity of 5 cities for 15 days

City A		*City B*		*City C*		*City D*		*City E*	
Max Temp	*Rel. Hum.*	*Max Temp*	*Rel Hum.*	*Max Temp*	*Rel Hum.*	*Max Temp*	*Rel. Hum.*	*Max Temp*	*Temp Hum.*
37.8	56.29	27.5	76.18	27.4	72.85	38.5	61.66	40.8	93.69
29.6	65.47	39.3	93.68	38.4	87.08	43.9	66 84	31.7	69.85
40.9	88.10	32.1	76.75	32.3	68.59	24.7	72.17	39.8	64.71
36 0	58.07	35.3	66.33	42.2	70.33	37.9	60.28	33.2	91.75
37.2	56.17	30.2	64.87	39.2	81.64	31.8	84.74	35.4	62.81
32.4	84.15	25.9	61.90	24.4	89.16	39.8	68.84	35.1	63.90
35.3	85.36	26.6	74.09	35.5	84.76	43.5	74.90	30.2	92.70
42.9	92.88	39.7	76.05	38.8	64.12	28.0	69.66	30.1	69.73
25.7	82.19	25.7	80.38	25.7	64.56	33.7	94.67	26.9	58.21
32.6	60.95	28.4	78.50	30.0	56 60	28.7	91.91	33.4	91.76
36.7	89.02	40.5	61.14	39.6	90.85	29.2	69.62	29.0	75.79
34.1	86.36	43.3	89.00	42.3	65.33	25.4	65.80	38.8	57.72
28.1	85.56	35.1	89.03	33.3	61.48	30.4	94.67	30.2	94.08
30.8	86.85	35.8	55.28	37.0	69.73	36.8	57.98	34.2	64.08
38.5	82.07	33.1	62.46	24.2	76.50	24.1	81.84	34.8	87.98
$y = 0.11x + 73.42$		$y = 0.4x + 60.4$		$y = 0.08 + 70.73$		$y = -0.62x + 94.84$		$y = -0.19x + 82.456$	

Note: x represents maximum temperature and y represents relative humidity.

66.3. Axes

X Axis: The axis is used to represent the sub-sample identification number.

Y Axis: This axis is used to represent the intercept of fitted least square lines.

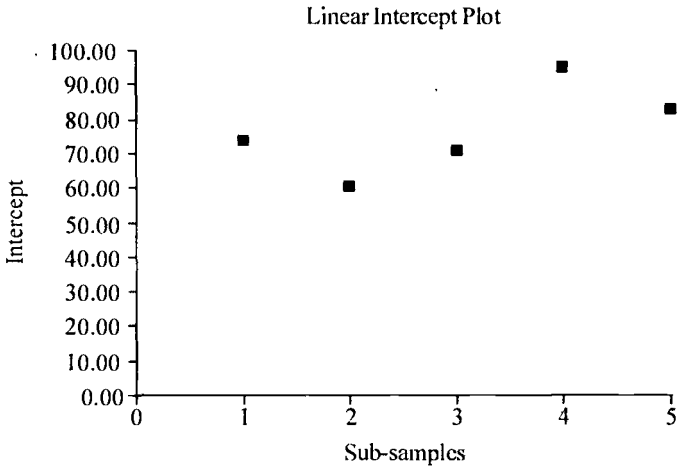


Fig. 66.1. *A Linear Intercept plot for the data provided in Table 66.1*

Here we see there is some shift in the value of the intercept over the different cities, but the difference may not be significant.

66.4. Uses

(*a*) The plot provides a means to compare the intercept of the least square lines fitted between two variables for various sub-samples.

(*b*) It helps in to study the relation between two variables when classified by another categorical variable.

66.5. Related Techniques

(*a*) Linear Slope Plot;

(*b*) Linear Residual Plot; and

(*c*) Linear Correlation Plot.

67. LINEAR SLOPE PLOT

67.1. Definition and Description

The linear slope plot is used to detect if the slope obtained from the least square linear fit between two variables differs across different sub samples. This can also be used to study variation in the slope between two variables across the different sub-samples under different categories. The plot is a scatter plot between the sub-sample identification number and the sub-sample slope obtained from the linear fit. In case of large variation in slopes the points will be scattered, otherwise they will form an approximate straight line parallel to X axis.

67.2. Working Data

The data for this purpose is same as that of Table 66.1.

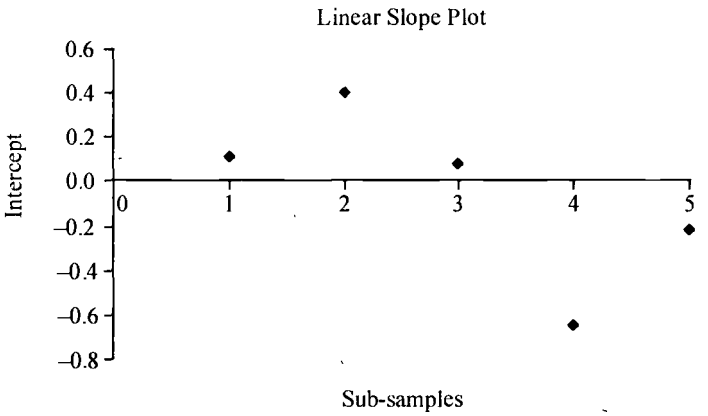


Fig. 67.1. *A Linear Slope plot for the data provided in Table 66.1*

Here we see there is slight variation in the value of the slopes over the different cities. Also in all cases the values of the slope is negligible and so the effect of maximum temperature is very less on relative humidity.

67.3. Axes

X Axis: The axis is used to represent the sub-sample identification number.

Y Axis: This axis is used to represent the slope of fitted least square lines.

67.4. Uses

(*a*) The plot provides a means to compare the slope of the least square lines fitted between two variables for various sub-samples.

(*b*) It helps us to study the relation between two variables when classified by another categorical variable.

(*c*) It helps to understand how much the extent of dependence of one variable on the other changes across the sub-samples.

67.5. Related Techniques

(*a*) Linear Intercept Plot;

(*b*) Linear Residual Plot; and

(*c*) Linear Correlation Plot.

68. LORENZ CURVE

68.1. Definition and Description

The Lorenz curve is a graphical representation of the cumulative percentage of the values of a variable. To build the Lorenz curve, all the elements of a distribution must be ordered from the maximum to the minimum. Then, each element is plotted according to their cumulative percentage of X and Y, X being the cumulative percentage of elements. For instance, out of a distribution of n elements of X namely $X_1, X_2, \ldots, X_n$ we perform the following calculations.

Serial	*Ordered values*	*Abscissa*	*Ordinates*
(i)	X_i	$x_i = \frac{i}{n} \times 100$	$y_i = \frac{\sum_{i=1}^{i} X_i}{\sum_{i=1}^{n} X_i} \times 100$

Accordingly we plot the points (x_i, y_i) and join them by a free hand smooth curve. The Lorenz curve is compared with the perfect equality line, which is a line that makes an angle of 45° with the positive direction of X axis and passes through the origin. This line is obtained only if each element has an equal value in its shares of x_i and y_i. If there is any inequality in the distribution of the variable, then the Lorenz curve falls below the line of equality. The total amount of inequality can be summarized by the Gini coefficient (also called the Gini ratio), which is the ratio between the area enclosed by the line of equality and the Lorenz curve, and the total triangular area under the line of equality. The more the value of the Gini's coefficient more is the inequality in the distribution of the variable.

68.2. Working Data

The data for this purpose consists of absolute values of a random sample drawn from a standard normal distribution. The necessary calculations for drawing the Lorenz curve are shown in Table 68.1 below.

Table 68.1: Data and calculations for Lorenz Curve

Serial no (i)	*Ordered values* X_i	*Abscissa* $x_i = \frac{i}{n} \times 100$	*Ordinates* $y_i = \frac{\sum_{i=1}^{i} X_i}{\sum_{i=1}^{n} X_i} \times 100$
1	0.1	5.6	0.63
2	0.1	11.1	0.127
3	0.2	16.7	2.53
4	0.2	22.2	3.8

Serial no. (i)	*Ordered values* X_i	*Abscissa* $x_i = \frac{i}{n} \times 100$	*Ordinates* $y_i = \frac{\sum_{i=1}^{i} X_i}{\sum_{i=1}^{n} X_i} \times 100$
5	0.2	27.8	5.06
6	0.3	33.3	6.96
7	0.4	38.9	9.49
8	0.6	44.4	13.29
9	0.6	50.0	17.09
10	0.7	55.6	21.52
11	0.8	61.1	26.58
12	1.0	66.7	32.91
13	1.2	72.2	40.51
14	1.6	77.8	49.37
15	1.8	83.3	60.76
16	1.8	88 9	72.15
17	2.2	94.4	86.1
18	2.2	100.0	100.0

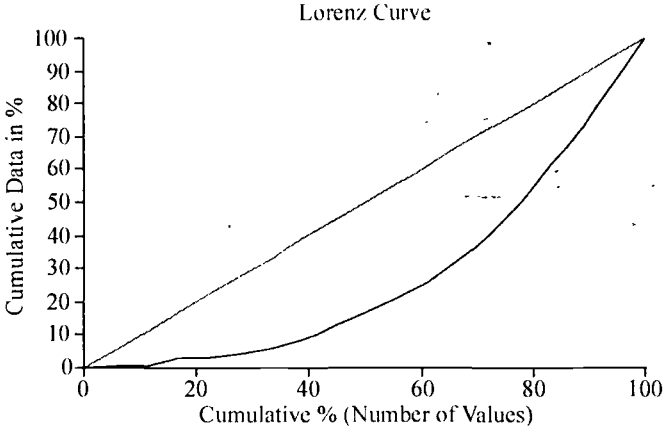


Fig. 68.1. *A Lorenz Curve for data in Table 68.1*

68.3. Axes

X Axis: The axis is used to represent the cumulative percentage of the number of observations.

Y Axis: This axis is used to represent the cumulative percentage of the response variable.

68.4. Uses

(*a*) The Lorenz curve is used in economics to describe inequality in the distribution of wealth amongst the population.

(*b*) The plot or the Gini's coefficient can be used to compare the pattern of distribution of wealth between countries.

68.5. Related Techniques

(*a*) Gini's Coefficient; and

(*b*) Lorenz Asymmetry Coefficient.

69. MOSAIC PLOT

69.1. Definition and Description

Mosaic plots are used for graphical representation of the two way or multi-way contingency tables. Mosaic plots can be stated as an extension of grouped bar-charts, where the width and height of the bars are proportional to the relative frequency of two variables. The mosaic display though appears to be a direct descendents of the bar charts, has features of sub-divided and proportional bars, introduced by Charles Joseph Minard in 1844. The modern form of mosaic display is attributed to Hartigan and Kleiner (1981).

The mosaic plot, proposed by Hartigan and Kleiner (1981) is basically a graphical method used to show the values of the contingency table which is cross classified by another categorical variable. In this diagram the cell frequency of the contingency table are represented by rectangular blocks or 'tiles' whose area is proportional to the cell frequencies provided in the table. Here each rectangle is drawn in such a way that the width of each rectangle is proportional to its column total and its height is proportional to the conditional frequency. Thus, the area of each tile is proportional to the cell frequency.

69.2. Working Data

The working data is provided in Table 69.1.The data is a two way contingency table derived from a 5-way contingency table used in Edwards and Kreiner (1983). The data come from a sample of employed men aged in between 18 and 67, who were asked whether, in the preceding year, they had carried out any work on their homes.

Table 69.1: A Contingency table showing age of respondent and labour type

	Age		
Labour type	*Less than 30*	*31 - 45*	*Greater than 46*
Skilled	169	123	68
Unskilled	161	128	160
Office	187	345	250

69.3. Interpretation of the Plot

It may be noted that the tiles are separated slightly, in order to make small counts more visible and to make better visualization of the tiles for understanding, horizontal and vertical dimensions. The particular display can also be used to study the independence between two contingency variables. If the tiles of each row have the same height then it indicates that the two categorical variables are independent. In such a case, all the rectangles in each row are aligned vertically.

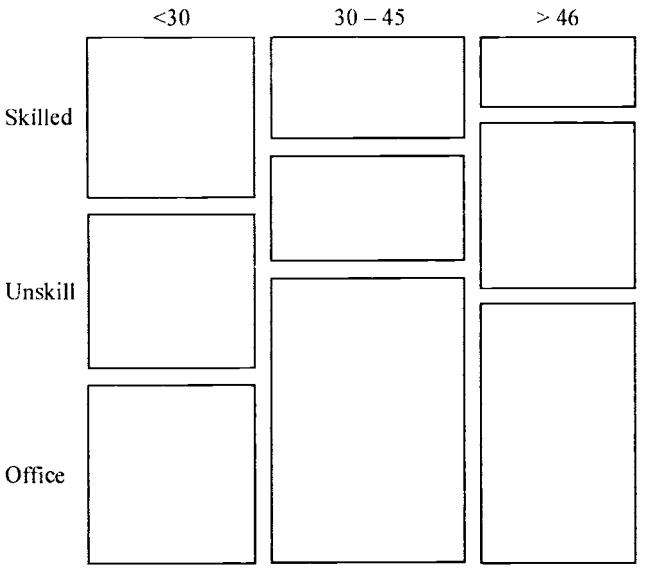


Fig. 69.1. *A mosiac display as proposed by Hartigan and Kleiner (1981)*

Here, r is the total number of rows and c is the total number of columns.

69.4. Advantages

(*a*) The plot is used for the visualization of data in case of data arranged in the form of contingency table.

(*b*) The plot can be used to understand the dependence if any between two different attributes.

(*c*) The plot is simple to draw and its interpretation is also easy.

69.5. Disadvantages

(*a*) The criterion of vertical alignment of rectangles remains difficult to be assessed visually especially in cases where the lack of alignments is not very loud.

(*b*) This display does not support any comparison between the observed and expected frequencies.

69.6. Friendly's Extension to Mosaic Display

Michael Friendly's works ushered a new era in the visual representation of categorical data. Friendly (1994) proposed some extension by the use of color which helps mosaic plots not only to display contingency tables but also visualizes the patterns of deviation from a specified model. Though the basic mosaic display proposed by Hartigan and Kleiner displays data in contingency tables but it does not in general provide a visual representation whether the data fits a specific model or not and also does not have any means to display the expected frequencies.

For a two-way contingency table, the expected frequency of each cell, *i.e.*, the frequency that a cell should have under the assumption that the two categorical variables (row and column variables) are independent can be easily computed. If e_{ij} denotes the expected frequency of the cell in the i^{th} row and j^{th} column, then we have

$$e_{ij} = \frac{r_i \times c_j}{n}$$

where, $r_i = i^{th}$ row total, $c_j = j^{th}$ column total and n = total frequency.

First the mosaic display is drawn based on the proposals made by Hartigan and Kleiner (1981) are then colored using a color scheme based on standardized (Pearsonian) residuals.

Let, d_{ij} be the standardized residual corresponding to the i^{th} row and j^{th} column. So, we have

$$d_{ij} = \frac{o_{ij} - e_{ij}}{\sqrt{e_{ij}}},$$

where $i = 1,2,\ldots,r$ and $j = 1, 2,\ldots, c$.

Here, r is the total number of rows and c is the total number of columns.

Now the value of d_{ij} is obtained for each cell. This value is used to determine the color to be provided to each of the rectangles. Under the color scheme used by Friendly, cells with positive deviations are colored blue and those with negative deviations are colored red. For $|d_{ij}| \leq 2$ the rectangles are empty and for $|d_{ij}| > 2$ the rectangles are filled, for $|d_{ij}| > 4$, the rectangles are filled with a darker pattern. For $|d_{ij}| \leq 1$ the dashed line is used to draw the boundary of rectangles and for $1 \leq |d_{ij}| \leq 2$ solid lines are used to draw the boundaries of the rectangles.

The color used for filling the rectangles or drawing the boundaries are determined by the sign of d_{ij}. The figure below gives the extended mosaic display by Friendly for the data given in Table 69.1.

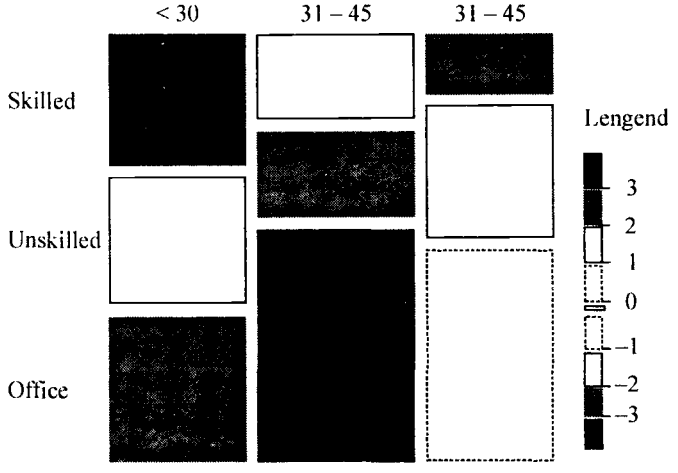


Fig. 69.2. *A Mosaic display with the extensions proposed by Friendly (1994)*

From the plot it appears that the two attributes are dependent on each other and there are 5 cells which are filled with darker patterns implying that in these cells the difference between observes cell frequencies and the expected cell frequencies are widely apart.

69.7. Related Techniques

(*a*) Chi-square test for independence in contingency table;
(*b*) Association Plot;
(*c*) Sieve Diagram; and
(*d*) Four Fold Display.

70. MOVING AVERAGE PLOT

70.1. Definition and Description

Moving average method is a simple device of reducing fluctuations and obtaining trend values with a fair degree of accuracy. When a trend is to be determined by this method, the average values for a number of years (months or weeks) is secured. Then this average is taken as the normal or trend value for the unit of time falling at the middle of the period covered in the calculation of the period. The averaged values are then plotted corresponding to the different years. The points are then connected by a free hand smooth curve. Thus time is considered along the X axis and corresponding moving average values along the Y axis. Some times the original time series is also plotted in the form of a line diagram so that a comparison can be made with the graph of moving average.

70.2. Working Data

The data used for this plot is a hypothetical one representing data related to the production of an industry for different years.

Table 70.1: Yearwise production of an industry in thousand tons.

Year	*Production '000 tons*	*Year*	*Production '000 tons*
1968	21	1973	22
1969	22	1974	25
1970	23	1975	26
1971	25	1976	27
1972	24	1977	26

Correspondingly we calculate 3-year moving averages, determine the trend of the time series. For calculating the 3-year moving averages we construct the following table:

Table 70.2: 3-Yearly moving averages based on data in Table 70.1

Year	*Production*	*3 Yearly Moving Total*	*3 Yearly Moving Average*
1968	21	—	—
1969	22	66	22
1970	23	70	23.33
1971	25	72	24
1972	24	71	23.67
1973	22	71	23.67

(contd...)

Year	*Production*	*3 Yearly Moving Total*	*3 Yearly Moving Average*
1974	25	73	24.33
1975	26	78	26
1976	27	79	26.33
1977	26	—	—

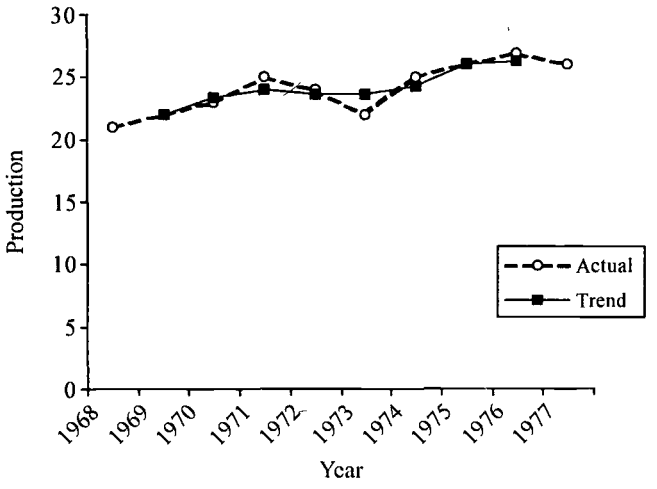


Fig. 70.1. *Moving Average Plot based on data in Table 70.2.*

70.3. Axes

X Axis: The axis is used to represent the time.

Y Axis: This axis is used to represent the values of the response variable. The variable is Production in this case.

70.4. Uses

(*a*) The plot provides a visualization of the trend value of the data.

(*b*) A comparison between the actual values and the trend values may also be made using this plot.

70.6. Related Techniques

(*a*) Fitting Trend Equation;

(*b*) Run Sequence Plot;

(*c*) Line Diagram; and

(*d*) Historigram.

71. MSE PLOT

71.1. Definition and Description

The MSE plot or the mean sum of square plot is a plot that can be used to locate the appropriate relation between a dependent variable and an independent variable. The plot can be used for time series data as well, to detect the most likely fitted model for a time series data from a host of given models. It can also be used to detect the same for a multiple number of time series based on the same time period.

Initially we fit a linear trend between the response variable with the independent variable. Then the MSE (mean sum of squares) is computed for the trend by taking the differences between the fitted value and the actual data by squaring them and averaging over all the values of the independent variable. Proceeding in this way we calculate the MSE for other trend equations like, quadratic, exponential and so on. The graph is then drawn by plotting the MSE values (along the vertical axis) corresponding to the different types of regression (along the X axis). Thus the X axis consists of the names of different regression equations like linear, quadratic, exponential etc. and the Y axis consisting of MSE values.

The graph can be drawn for several time series data as well.

71.2. Working Data

The table given below is taken from the Economic Survey, 2001-02 and shows the gross national product and net national product of India.

Table 71.1: GNP and NNP of India for selected years

Year	*GNP at factor cost (Rs. crores)*	*NNP at factor cost (Rs crores)*	*Per Capita NNP (Rs.)*
1951	5.7	5.4	3.6
1961	5.4	5	2.6
1971	6.4	6.1	3.6
1981	16.5	16.2	14
1982	11	10.4	7.9
1983	17	17.4	15
1984	11.9	11.6	9.2
1985	12.1	11.4	9
1986	11.4	11.1	8.8
1987	13.4	13.5	11.1
1988	19.3	19.6	17.1
1989	15.6	15.4	13

(contd...)

Year	*GNP at factor cost (Rs crores)*	*NNP at factor cost (Rs. crores)*	*Per Capita NNP (Rs.)*
1990	16.5	16.7	14.3
1991	15	14.3	12
1992	14.3	14.1	12
1993	16.3	16.8	14.2
1994	17.5	17.5	15.2
1995	17.2	16.9	14.6
1996	16.1	16 1	13.9
1997	11.9	12	9.9
1998	15	15.5	13.3
1999	9.9	10.1	8.1
2000	7.9	7.8	5.9

Table 71.2: MSE for different types of regression based on data in Table 71.1

Fitted equation	*GNP*	*NNP*	*Per Capita* NNP
Linear	11.3353	11.9716	11.3569
Quadratic	9 8508	10.6929	10.5625
Logarithmic	11.3097	11.9439	11.3569
Exponential	12.4115	13.1914	13.1044
Power curve	12.8164	13.69	13.7641

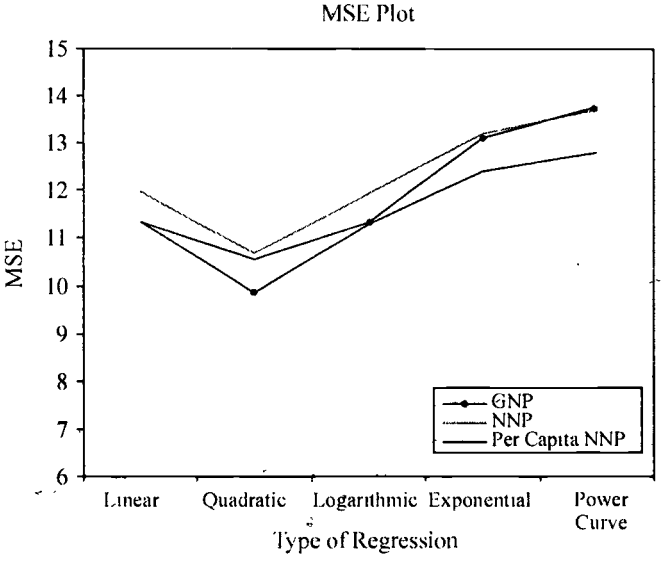


Fig. 71.1. *An MSE Plot for the data in Table 71.2*

From the MSE curves we find that for all the time series the quadratic equation is the best fit. Also the pattern of the MSE curves are same for all the time series.

71.3. Axes

X Axis: The axis consists of the names of different regression equations like linear, quadratic, exponential etc.

Y Axis: In the vertical axis we consider the corresponding MSE values of the trend equations are considered.

71.4. Advantages

(*a*) To visualize the MSE values for different types of regression equations of the same time series.

(*b*) To compare the distributional pattern of several time series data.

(*c*) Provides a tool for searching the best model for a given time series data.

71.5. Disadvantages

(a) Most of the commonly used statistical software does not provide the option to draw the plot.

(b) The calculations involved are very lengthy.

(c) The significant difference between the MSEs computed for different time series data cannot be computed.

71.6. Some Related Techniques

(*a*) Mean Sum of Squares;

(*b*) Tests for Goodness of Fit; and

(*c*) Test for Linearity of Regression.

72. NORMAL PROBABILITY PLOT

72.1. Definition and Description

A substantial portion of inferential Statistics is based on the assumption that the underlying distribution from which the sample is drawn follows a normal probability distribution. The normal probability plot is the plot used for checking the normality of a data set. The normality assumption acts as a gatekeeper to many statistical methodologies relating to estimation, testing of hypothesis and the like. A quick way to assess the normality assumption can be performed by the normal probability plot.

Let $y_1, y_2, \ldots y_n$ be a random sample of size n. We are to test if the random sample comes from a normal distribution. The steps for the construction of a probability plot are as follows:

(*a*) Sample is first ordered or arranged in increasing order of magnitude. The arranged sample is thus denoted by $y_{(1)}, y_{(2)}, \ldots y_{(n)}$

(*b*) For $i = 1,2,\ldots,n$ the values of $(i-0.5)/n$ are calculated.

(*c*) We then find $z_i = \phi^{-1}((i-0.5)/n)$, where ϕ denotes the standard normal distribution function.

(*d*) The points $(y_{(i)}, z_i)$, $i = 1, 2,\ldots,n$ are then plotted in a graph sheet. If the plotted points fall approximately along the straight line then it may be concluded that the random sample comes from a normal distribution, otherwise not.

72.2. Working Data and Calculations

The data for this purpose represents the head length of 25 sons, originally taken from Anderson (1958).

Table 72.1: Data and calculations for normal probability plot

$y_{(i)}$	$(i-0.5)/n$	$z_i=\phi^{-1}((i-0.5)/n)$
163	0.0208	−2.0368
174	0.0625	−1.5341
174	0.1041	−1.2581
175	0.1458	−1.0544
176	0.1875	−0.8871
176	0.2291	−0.7415
179	0.2708	−0.6102
181	0.3125	−0.4887
181	0.3541	−0.3741

(contd...)

$y_{(i)}$	$(i-0.5)/n$	$z_i=\phi^{-1}((i-0.5)/n)$
183	0.3958	–0.2641
183	0.4375	–0.1573
186	0.4791	–0.0522
188	0.5208	0.0522
188	0.5625	0.1573
189	0.6041	0.2641
190	0.6458	0.3740
191	0.6875	0 4887
192	0.7291	0.6102
192	0.7708	0.7415
195	0.8125	0 8871
195	0.8541	1.0544
197	0.8958	1.2581
197	0.9375	1.5341
208	0.9791	2.0368

72.3. A Test to Check the Normality of Data

The normal probability plot that is based on the data shows that the head lengths of the sons follow normal distribution. The result can also be conformed by using Anderson-Darling test. The Anderson-Darling test is a modification of the K-S test as it gives more weight to the tails than the K-S test. It was also shown by Stephen (1974), that for a wide class of alternatives with unknown parameters, that Anderson-Darling test is typically more powerful than any other distance test for testing normality.

ANDERSON-DARLING 1 -SAMPLE TEST
THAT THE DATA CAME FRM A NORMAL DISTRIBUTION

1. STATISTICS:
   - NUMBER OF OBSERVATIONS = 25
   - MEAN = 185.5417
   - STANDARD DEVIATION = 9.930099
   - ANDERSON-DARLING TEST STATISTIC VALUE = 0.2348938
   - ADJUSTED TEST STATISTIC VALUE = 0.2638477
2. CRITICAL VALUES:
   - 95 % POINT = 0.7350000
   - 99 % POINT = 1.021000
3. CONCLUSION (AT THE 5% LEVEL):
   - THE DATA COME FROM A NORMAL DISTRIBUTION.

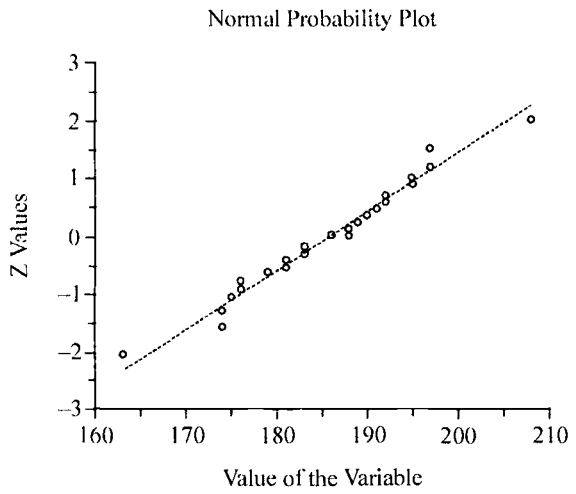


Fig. 72.1. *A normal probability plot to the data in Table 72.1*

On looking at the Figure 72.1 it appears that the data is a good fit to normal distribution.

72.4. Axes

X Axis: The axis is used to represent the values of the observations.

Y Axis: This axis is used to represent the Z values corresponding to each value of the observation.

72.5. Advantages

(*a*) Normal Probability plots provide a quick check to normality of the data, which is often a necessary condition for many statistical tests.

(*b*) In regression or ANOVA, we often assume that the residual component is normally distributed. This plot can be used to check the validity of that.

72.6. Disadvantage

The problem with this type of plot is that the conclusions regarding the goodness of fit test taken by different users may be subjective. Since, no proper confidence bands are developed around the straight line obtained from the plotting of expected frequency from the normal distribution, so accessing the goodness of fit, may sometimes become difficult even for the experts.

72.7. Related Techniques

(*a*) Chi-square test for the goodness of fit;

(*b*) Anderson-Darling test;

(*c*) Wilks Shapiro test;

(*d*) EDF plot; and

(*e*) Probability Plot.

73. *np* CHART

73.1. Definition and Description

np- Chart is a type of control chart used for attributes and is called as the control chart for number of defectives. Control charts are used to study the variability in the quality of a manufactured product and are used to understand if the process is within control. If the quality is not directly measurable but can be classified as defective or non-defective then this type of control chart can be used.

Here k samples of the manufactured product are collected of same size 'n' (say). Let d_i be the number of defects in the i^{th} ($i = 1, 2, \ldots, k$) sample. For a quality product the number of defective items in a sample of size n follows binomial distribution. If the number of defective items in the population (np) is not known then it can be estimated by $n\bar{p}$ where

$$\bar{p} = \frac{1}{k}\sum p_i = \frac{1}{k}\sum \frac{d_i}{n}$$

A control chart consists of three lines parallel to the X axis, the control line, the upper control limit (UCL) and the lower control limit (LCL). Here we have,

Control Line = $n\bar{p}$

Upper Control Limit = UCL = $n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})}$

Lower Control Limit = LCL = $n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})}$

After all this calculations are done the values of d_i are computed for each of the samples and then they are plotted for different values of i. In other words the values of the defectives in the different samples are plotted in the form of dots against the sample numbers. The UCL and LCL are also drawn. If all the dots *i.e.*, (i, d_i) fall within the UCL and LCL (control limits) then the system is said to be under control otherwise the system is beyond control.

73.2. Working Data

The data for this purpose is taken from a voltage stabilizer manufacturing company. The quality control inspector of the company tests the quality of 50 units of its product daily for 15 days and accordingly finds following defective items. The fraction defective are also computed in the table drawn below:

Table 73.1: Number of defectives in different days

Days	*No of Defective*	*Days*	*No of Defective*
1	5	9	1
2	10	10	8
3	3	11	6
4	2	12	7
5	8	13	4
6	1	14	5
7	4	15	3
8	3	—	—

Here, $\bar{p} = \frac{\sum p_i}{k} = \frac{1.4}{15} = 0.093$ and $n = 50$.

Upper Control Limit = $UCL = n\bar{p} + 3\sqrt{n\bar{p}(1-\bar{p})} = 4.65 + 3\sqrt{4.65 \times 0.907} = 10.81$

Lower Control Limit = $LCL = n\bar{p} - 3\sqrt{n\bar{p}(1-\bar{p})} = 4.65 - 3\sqrt{4.65 \times 0.907} = -$ ve values.

So, LCL = 0

Control Limit = CL = $n\bar{p} = 4.65$.

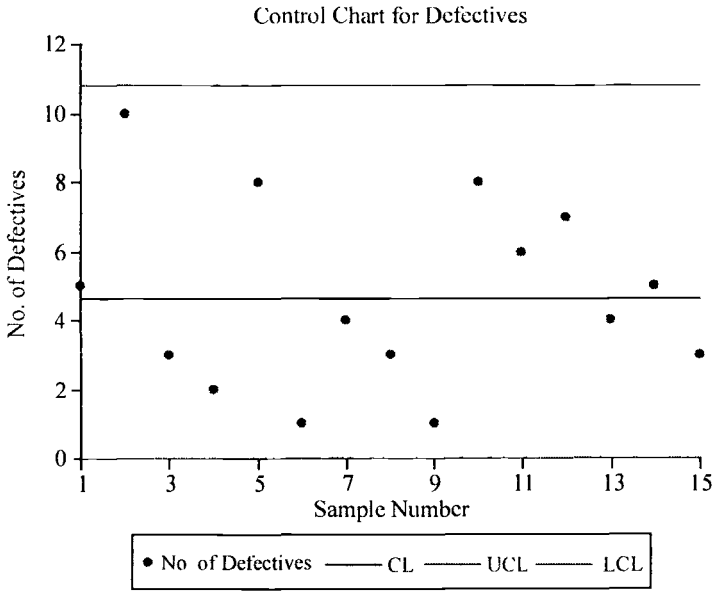


Fig. 73.1. *np chart for the data in Table 73.1*

The control chart shows that the number of defective in none of the samples is outside the control limits and accordingly we may comment that the system is within control.

73.3. Axes

X Axis: The axis is used to represent the sample numbers.

Y Axis: This axis is used to represent the values of the number of defectives in the sample.

73.4. Uses

(*a*) The plot is simple to draw and the calculations are relatively simple.

(*b*) The chart is used to discover the existence of any assignable cause of variation.

(*c*) The chart is successful even in case of small samples.

(*d*) The chart can deal with situations where the quality of the product cannot be measured but is only an attribute.

73.5. Related Techniques

(*a*) Process Control;

(*b*) Control Chart;

(*c*) σ Chart;

(*d*) X Chart; and

(*e*) p Chart.

74. Ogive

74.1. Definition and Description

The graphical representation of cumulative frequency distribution for a continuous variable is called the cumulative frequency polygon or ogive. The cumulative frequency curve can be of two types either more-than type or less-than type. In drawing a less than type ogive the upper class boundaries (and not class limits) are plotted along the X-axis and the corresponding less than type cumulative frequencies are plotted along the Y-axis. For drawing a more than type ogive similar process is followed but the lower class boundaries (and not class limits) are considered. The points thus obtained are joined by a free hand smooth curve and the resulting curve is called as the cumulative frequency curve or ogive.

The ogive of less than type starts from the lowest class boundary on the horizontal axis gradually rising upward and ending at the highest class boundary corresponding to the cumulative frequency which is equal to total frequency. It looks like an elongated 'S'. The more than type ogive has the appearance of an elongated 'S' turned upside down. The more than type ogive and the less than type ogive intersects at a point, the x coordinate of that point is the median of the frequency distribution.

74.2. Working Data

The working data is a hypothetical frequency distribution showing the marks of students along with their corresponding frequencies and cumulative frequencies.

Table 74.1: Frequency distribution of the marks of 95 students

Class	*Frequency*	*Cumulative Frequency*
0–10	5	5
10–20	4	9
20–30	8	17
30–40	12	29
40–50	16	45
50–60	25	70
60–70	10	80
70–80	8	88
80–90	5	93
90–100	2	95

74.3. Axes

X Axis: It can represent the upper limits of class intervals.

Y Axis: The vertical axis is used for representing the cumulative frequencies. For this case we took the number of students along the y axis.

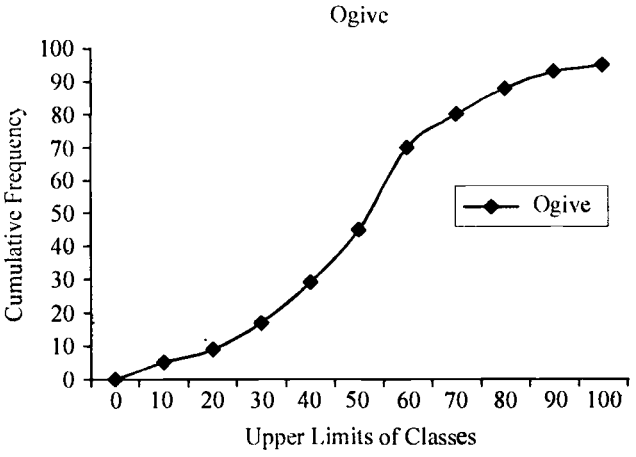


Fig. 74.1. *The less than type Ogive for the data in Table 74.2*

74.4. Advantages

(*a*) The ogive is used to find the median, quartiles, deciles and percentiles of a frequency distribution.

(*b*) It is also useful in finding the cumulative frequency corresponding to a given value of the variable.

(*c*) The curve can also be used to find the number of observations, which are expected to lie between two values of the variable.

74.5. Disadvantages

(*a*) The curve cannot be drawn for open-end classes.

(*b*) Though the plot is very common yet most commonly used statistical software provides the option of drawing the plot.

74.6. Related Techniques

(*a*) Frequency Polygon;

(*b*) Histogram; and

(*c*) Frequency Distribution.

75. Ord Plot

75.1. Definition and Description

Ord (1967) suggested a simple plot that can be used as a tool to diagnose the form of a discrete distribution. Ord showed that for a number of discrete distribution like the Poisson, binomial, negative binomial, logarithmic series distribution etc. the expression

$$\frac{xP[X=x]}{P[X=x-1]} = a + bx$$

i.e., of the linear form. Also the slope and intercept varies from distribution to distribution.

For example, in case of Poisson distribution we have,

$$P(X=x) = \frac{e^{-\lambda}\lambda^x}{x!},\ \lambda > 0, x = 0, 1, 2, \ldots$$

$$\Rightarrow \quad \frac{P[X=x]}{P[X=x-1]} = \frac{e^{-\lambda}\lambda^x}{x!} \Big/ \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \frac{\lambda}{x}$$

Thus, $$\frac{xP[X=x]}{P[X=x-1]} = \lambda \qquad \ldots(75.1)$$

In case of binomial distribution, we have

$$P[X=x] = {}^nC_x p^x (1-p)^{n-x}, x = 0, 1, 2, \ldots, n$$

$$\Rightarrow \quad \frac{P[X=x]}{P[X=x-1]} = {}^nC_x p^x (1-p)^{n-x} \Big/ {}^nC_{x-1} p^{x-1} (1-p)^{n-x+1}$$

$$= \frac{n-x+1}{x}\frac{p}{1-p}$$

Thus, $$\frac{xP[X=x]}{P[X=x-1]} = \frac{(n+1)p}{1-p} - \frac{p}{1-p}x \qquad \ldots(75.2)$$

In case of negative binomial distribution, we have

$$P[X=x] = {}^{x+r-1}C_x p^r q^x,\ x = 0, 1, 2, \ldots; p + q = 1$$

$$\Rightarrow \quad \frac{P[X=x]}{P[X=x-1]} = {}^{x+r-1}C_x p^r q^x \Big/ {}^{x-1+r-1}C_{x-1} p^r q^{x-1} = \frac{x+r-1}{x}q$$

Thus, $$\frac{xP[X=x]}{P[X=x-1]} = (r-1)q + qx \qquad \ldots(75.3)$$

In case of the logarithmic series distribution we have,

$$P(X=x) = -\frac{\theta^x}{x\log(1-\theta)},\ 0 < \theta < 1, x = 0, 1, 2, \ldots$$

$$\Rightarrow \quad \frac{P[X=x]}{P[X=x-1]} = \frac{-\theta^{x}}{x\log(1-\theta)} \times \frac{(x-1)\log(1-\theta)}{-\theta^{x-1}} = \frac{x-1}{x}\theta$$

Thus, $$\frac{xP[X=x]}{P[X=x-1]} = -\theta + \theta x \qquad ...(75.4)$$

In case of the geometric distribution we have,

$$P(X=x) = q^{x},\ 0<p<1, x=0,1,2,\ldots$$

$$\Rightarrow \quad \frac{P[X=x]}{P[X=x-1]} = \frac{q^{x}p}{q^{x-1}p} = q$$

Thus, $$\frac{xP[X=x]}{P[X=x-1]} = qx \qquad ...(75.5)$$

In case of the discrete uniform distribution we have,

$$P(X=x) = \frac{1}{n+1}\ x=0,1,2,\ldots,n$$

$$\Rightarrow \quad \frac{P[X=x]}{P[X=x-1]} = \frac{1/(n+1)}{1/(n+1)} = 1$$

Thus, $$\frac{xP[X=x]}{P[X=x-1]} = x \qquad ...(75.6)$$

So, we find that from each of the equations (75.1), (75.2), (75.3), (75.4), (75.5) and (75.6) the form of the expression is a straight line but each of the straight line has a different form compared to the other in terms of their slopes and intercepts. The following table can explain it more clearly:

Table 75.1: Clue to the understanding of the type of distribution from Ord Plots

Distribution	*Slope*	*Intercept*	*Estimate of the parameter*
Poisson (λ)	0	+	λ = *intercept*
Binomial (*n, p*)	–	+	*p = slope/(slope - 1)*
Neg. binomial *(n, p)*	+	+	*p = 1 – slope*
Logaritmic Series distribution	(q)	+	*– θ = slope*
Geometric *(p)*	+	0	*p = 1 – slope*
Discrete uniform	+1	0	

But in cases where each observation cannot be treated equally one may think of using the weighted least square method to maximize the efficiency of parameter estimation. Friendly (2003) advocated the use of weighted least squares to fit a line with $\frac{xP[X=x]}{P[X=x-1]}$ as the dependent variable and x as the independent variable. According to Friendly, the weights $w_x = \sqrt{o_x - 1}$ provides a reasonably good automatic diagnosis of the form of a probability distribution.

Actually, the weighted least square method has some advantages over the ordinary least square technique. They are:

(*i*) The weighted least square is an efficient method even for small data sets.

(*ii*) It has the ability to handle regression situations in which the data points are varying in quality.

Keeping these two points in mind Friendly must have advocated in favour of weighted least square instead of ordinary least squares.

Here, let

$$y_x = \frac{xP[X=x]}{P[X=x-1]} \text{ and } w_x = \sqrt{o_x - 1} \qquad ...(75.6)$$

So, we are to fit the line of the form, $y_x = a + b\,x$, where a and b are constants to be determined in such a way that,

$$S = \sum_{\lambda=1}^{\sigma} w_\lambda \left[y_\lambda - (a+bx)\right]^2 \qquad ...(75.7)$$

where a and b are constants to be determined in such a way that the error sum of squares *i.e.*, S is minimum. The values of a and b are determined with the help of two normal equations $\frac{\partial S}{\partial a} = 0$ and $\frac{\partial S}{\partial b} = 0$.

Now

$$\frac{\partial S}{\partial a} = 0 \Rightarrow \sum_{\lambda=1}^{\infty} w_\lambda y_\lambda = a\sum_{x=1}^{\infty} w_x + b\sum_{\tau=1}^{\infty} w_\lambda x \qquad ...(75.8)$$

Also

$$\frac{\partial S}{\partial b_j} = 0 \Rightarrow \sum_{\tau=1}^{\infty} w_\lambda y_\lambda x = a\sum_{\lambda=1}^{\infty} w_x x + b\sum_{x=1}^{\infty} w_x x^2 \qquad ...(75.9)$$

Solving (75.8) and (75.9) we can find the values of a and b and accordingly the weighted least square line is obtained.

75.2. Working Data

The data used for drawing the plot is taken from Jeffers (1978) which corresponds to the number of notices per day with its corresponding frequencies. The data is provided in Table 55.1. Table 75.2 shows the calculations required for drawing the plot and deriving the fitted lines (weighted and unweighted).

Table 75.2: Calculations on data from Jeffres (1978) for Ord Plot

x	o_x	$P(X=x)=\frac{o_x}{\sum o_x}$	$y_x=\frac{xP[X=x]}{P[X=x-1]}$	$w_x=\sqrt{o_x-1}$	*Estimated values of Y*	
					Least squares	*Weighted Least Squares*
0	162	0.173448			1.6819	1.5872
1	267	0.285867	1.648148	16.30951	1.8253	1.7636
2	271	0.29015	2.029963	16.43168	1.9687	1.94
3	185	0.198073	2.04797	13.56466	2.1121	2.1164
4	111	0.118844	2.4	10.48809	2.2555	2.2928
5	61	0.06531	2 747748	7.745967	2.3989	2.4692
6	27	0.028908	2.655738	5.09902	2.5423	2.6456
7	8	0.008565	2.074074	2.645751	2.6857	2.822
8	3	0.003212	3	1.414214	2.8291	2.9984
9	1	0.001071	3	0	2.9725	3.1748

Based on calculations in the above table we have

$$\sum_x w_x y_x = 157.74,\ \sum_x w_x = 73.698,\ \sum_x w_x x = 230.977,\ \sum_x w_x y_x x = 537.64,\ \text{and}$$

$$\sum_x w_x x^2 = 969.29$$

So, the weighted least square line using equations (75.8) and (75.9) is found to be,

$$y = 1.5872 + 0.1764x \qquad ...(75.10)$$

Also, using the values

$$\sum y = 21.60364,\ \sum x = 45, \sum xy = 116.64366 \text{ and } \sum x^2 = 285$$

we solve the normal equations and accordingly get the least square line as,

$$y = 1.6819 + 0.1434x \qquad ...(75.11)$$

75.3. Axes

X Axis: It represents the value of the independent variable. In the figure we take the number of notices per day is considered along the axis.

Y Axis: The vertical axis consists of the values of the expression $y_x = \frac{xP[X=x]}{P[X=x-1]}$.

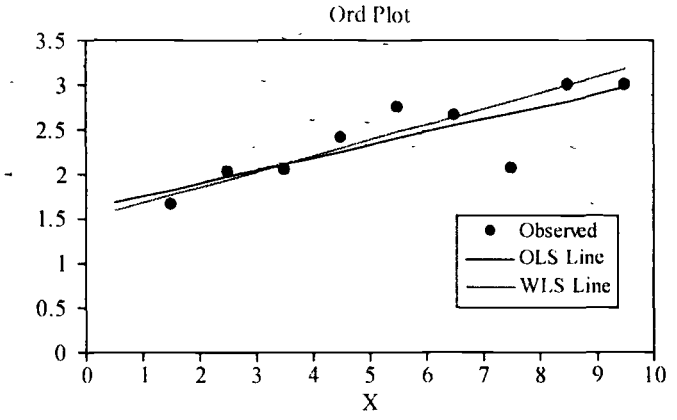


Fig. 75.1. *An Ord Plot for the data from Jeffres (1978) along with the fitted lines*

The plot shows that the data can be considered to have come from a Poisson distribution. Both the lines WLS and OLS lines provide a reasonably good fit for the plots. The intercept of the OLS line is 1.5872 and that of WLS line is 1.6819 which are the estimates of λ *under the OLS and WLS techniques respectively.*

75.4. Advantages

(*a*) The plot can be used as a diagnostic tool for the distribution that best fits the data.

(*b*) The slope or intercept of the fitted line acts as estimates of the parameters.

(*c*) The plot puts forward a visual impact on the quality of fit.

(*d*) The plot can be used to understand the pattern of variation between the observed and expected frequencies. For example, in the above figure one would find that the maximum variation for $X = 7$. Ordinary goodness of fit tests does not reveal such facts.

75.5. Disadvantages

(*a*) Though the plot provides a visualization of the difference between the type of distribution but in most cases it is difficult to infer with certainty about the type of fit.

(*b*) The graph requires a lot of computation and the calculations are more than the chi-square test for goodness of fit.

75.6. Related Techniques

(*a*) Chi-square Goodness of Fit Test;

(*b*) Rootogram;

(*c*) Chigram;

(*d*) Poissonness Plot; and

(*e*) Dubey Plot.

76. Parallel Co-ordinate Plot

76.1. Definition and Description

The parallel coordinate is a technique used for the representation of multivariate data. It was introduced by Inselberg (1985). The plot consists of a number of axes parallel to each other generally placed vertically. Each axis represents a variable. Thus if there are p-variates in the system then we have p axes each parallel to the other. The maximum and minimum values of a particular dimension are usually scaled to the upper and lower points of these vertical lines. A multivariate observation is been represented by p-1 line segments connecting each of the vertical axes at appropriate dimensional value. Thus a particular observation will be replaced by a zig-zag line running across all the parallel axes.

In the parallel coordinate plot, a p-dimensional point $(x_1, x_2, \ldots, x_p)$ is represented by a zig-zag line or a curve running across the plot. Speaking a little more elaborately for plotting a p-dimensional point $(x_1, x_2, \ldots, x_p)$ we have p-parallel axes with x_1 being plotted on axis 1, x_2 on axis 2 and so on through x_p on axis p The points plotted in this way are then connected using $(p - 1)$ straight lines. Thus in case of the plot any number of response variables can be represented with varying scales of measurement.

76.2. Working Data

The working data is a hypothetical table consisting of 10 variables. There are 11 observations in all.

Table 76.1: A hypothetical data of 10 multiple variables

	X_1	X_2	X_3	X_4	X_5	X_6	X_7	X_8	X_9	X_{10}
Case1	22	56	56	54	33	2	32	32	21	21
Case 2	78	43	73	34	44	3	43	34	32	43
Case 3	98	46	42	23	23	4	65	57	43	56
Case 4	54	57	34.6	56	31	5	76	89	56	78
Case5	43	89	78.9	8	24	6	43	32	78	94
Case 6	45	56	54	9	3	7	54	21	9	3
Case 7	45	34	23	32	65	43	32	32	43	35
Case 8	67	56	55	35	54	32	45	34	53	32
Case 9	89	78	54	67	46	12	67	56	42	15
Case 10	23	90	54	89	57	35	89	78	32	46
Case 11	41	21	33	78	78	67	32	90	45	45

76.3. Axes

In parallel co-ordinate plot we have only vertical axes and the number of axes are equal to the number of variables. The axes are parallel to each other and each one is dedicated for the representation of a particular variable.

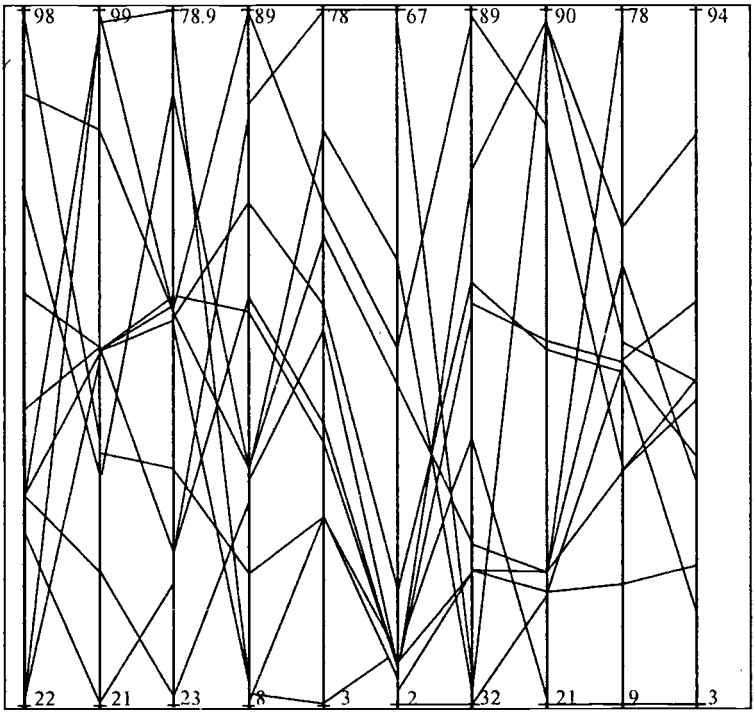


Fig. 76.1. *A parallel coordinate plot for data in Table 76.1*

76.4. Advantages

(*a*) The biggest advantage of the plot is definitely to make a representation of the multivariate data in a two dimensional plane.

(*b*) If we study one particular axis separately then we can easily detect the concentration of the points and hence can understand the central value of the variate.

(*c*) The plot can be used for subdividing the multivariate observations into groups or clusters.

(*d*) The graph can be used to identify the extent of linear relation between variables represented by adjacent axes.

(*e*) It will be easier to detect the highest order statistics as well as the lowest order statistics for each variable as well as for the multivariate data.

(*f*) The mode can be easily identified. The central value of the most intense concentration of the observations will be the mode vector.

76.5. Disadvantage

(*a*) The use of single color for the display makes the individual observation loose its identity even for a moderately large sample size.

(*b*) For a large sample size the line segments creates such over-plotting that the nature of the points can be understood only in an aggregate and not individually. Even by using different colors for different observations the problem of over-plotting may not be overcome properly.

76.6. Related Techniques

(*a*) Andrews Curve; and

(*b*) Icon Plots.

77. PARETO PLOT

77.1. Definition and Description

Pareto (1897) observed that in many countries the number of individuals in the population whose income exceeded a given level x was well approximated by $Cx^{-\alpha}$ for some real number C and $\alpha > 0$. Thus if $f(x)$ represents the number of individuals in the population with income more than x then we have,

$$y = Cx^{-\alpha} \qquad \ldots(77.1)$$

If we take logarithm of both sides then we have,

$$\log y = \log C - \alpha \log x$$

$$\Rightarrow \qquad Y = A - \alpha X,$$

Where $Y = \log y$, $A = \log C$ and $X = \log x$

Thus, the data of income (x) and population with income more than x (y), when plotted in a double logarithmic graph paper will approximately follow a straight line with slope $-\alpha$. However, for getting the curve we are to plot income along the X-axis and population with income more than x along the Y-axis. The Pareto curve for most of the population is extremely asymmetrical like that of a hyperbola.

77.2. Working Data

The data set for this purpose is a hypothetical figure which shows the distribution of income of urban households of a country in percentage.

Table 77.1: Distribution of income of urban households of a country in percentage

Income Class (Rs)	*Percentage of households*	*Income Class (Rs)*	*Percentage of households*
Under 500	13.6	5000–5999	1.7
500–999	28.9	6000–7999	1.7
1000–1999	32.5	8000–9999	0.7
2000–2999	10.6	10000–14999	0.8
3000–3999	5.6	15000–24999	0.5
4000–4999	3.1	25000 and above	0.3

Table 77.2: Necessary calculations for Pareto Curve and log transformations

Income (x)	*Percentage of households having income more than x (y)*	*log (x)*	*log (y)*
0	100		2
500	86.4	2.69897	1.936514
1000	57.5	3	1.759668
2000	25	3.30103	1.39794
3000	14.4	3.477121	1.158362
4000	8.8	3.60206	0.944483
5000	5.7	3.69897	0.755875
6000	4	3.778151	0.60206
8000	2.3	3.90309	0.361728
10000	1.6	4	0.20412
15000	0.8	4.176091	–0.09691
25000	0.3	4.39794	–0.52288

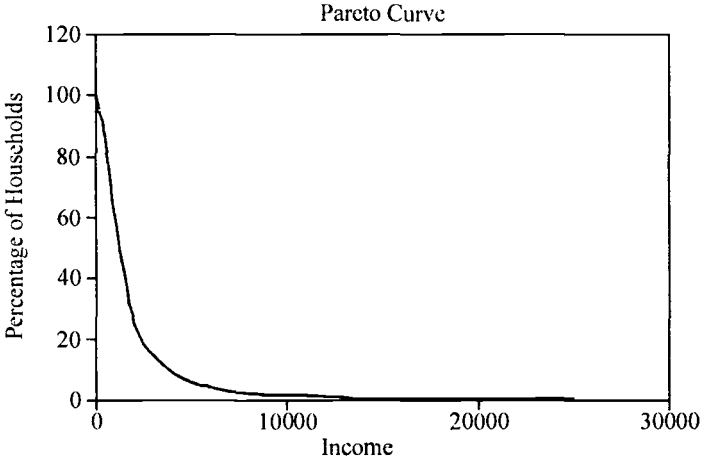


Fig. 77.1. *A Pareto curve for data provided in Table 77.1*

77.3. Axes

X Axis: In the Pareto curve the axis is used to represent the income.

Y Axis: The vertical axis we consider the number of individuals having income more than x. In the above example as the data is in percentage so percentages are plotted along the y-axis

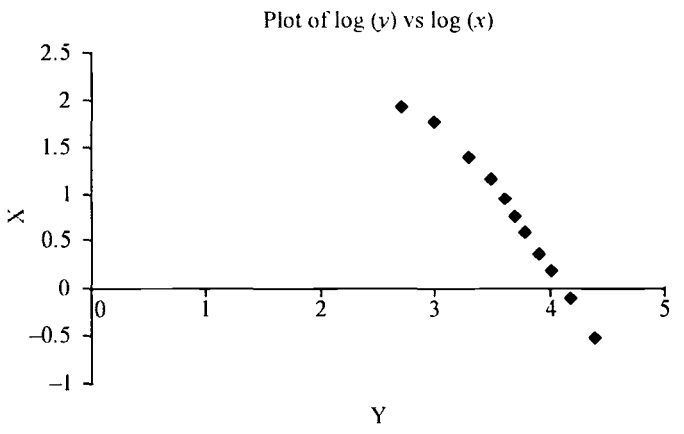


Fig. 77.2. *Plotting the values of x and y after logarithm transformation*

From Figure 77.2 we see that the plotted points are approximately linear. Which is an indication that the data abides by Pareto's law.

77.4. Uses

(*a*) To visualize the distribution of income of a country.

(*b*) The Pareto's law holds good for a capitalist economy. So this plot gives us an idea about the type of economy in the country.

(*c*) The plot can be used to find out the number of people which has income more than a particular amount.

77.5. Some Related Techniques

(*a*) Engel's Law;

(*b*) Lorenz Curve;

(*c*) Pareto's Law of Income Distribution.

78. *p* CHART

78.1. Definition and Description

p-Chart is a type of control chart used for attributes and is called as the control chart for fraction defective.Control charts are used to study the variability in the quality of a manufactured product and are used to understand if the process is within control. If the quality is not directly measurable but can be classified as defective or non-defective then this type of control chart can be used.

Here k samples of the manufactured product are collected, of same size 'n' (say). Let x_i be the number of defects in the i^{th} ($i = 1, 2, \ldots, k$) sample. For a quality product the number of defective items in a sample of size n follows binomial distribution. If the population fraction defective (p) is not known then it can be estimated by $\bar{p}$ where

$$\bar{p} = \frac{1}{k}\sum p_i = \frac{1}{k}\sum \frac{x_i}{n}$$

A control chart consists of three lines parallel to the X axis, the control line, the upper control limit (UCL) and the lower control limit (LCL). Here we have,

Control Line = $CL = \bar{p}$

Upper Control Limit = $UCL = \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$ and Lower Control Limit = $LCL = \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}}$

After all this calculations are done the values of p_i are computed for each of the samples and then they are plotted for different values of i. In other words the values of the fraction defective in the different samples are plotted in the form of dots against the sample numbers. The UCL and LCL are also drawn. If all the dots, *i.e.*, (i, p_i) fall within the UCL and LCL (control limits) then the system is said to be under control otherwise the system is beyond control.

78.2. Working Data

The data for this purpose is taken from a voltage stabilizer manufacturing company. The quality control inspector of the company tests the quality of 50 units of its product daily for 15 days and accordingly finds following defective items. The fraction defective are also computed in the table drawn below:

Table 78.1: Number of defectives in different days

Days	*No. of defective*	*Fraction defective*	*Days*	*No. of defective*	*Fraction defective*
1	5	0.1	9	1	0.02
2	10	0.2	10	8	0.16
3	3	0.06	11	6	0.12
4	2	0.04	12	7	0.14
5	8	0.16	13	4	0.08
6	1	0.02	14	5	0.1
7	4	0.08	15	6	0.06
8	3	0.06			

Here,

$$\bar{p} = \frac{\sum p_i}{k} = \frac{1.4}{15} = 0.093$$

$$\text{Upper Control Limit} = \text{UCL} = \bar{p} + 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.093 + 3\sqrt{\frac{0.093 \times 0.907}{50}} = 0.216$$

$$\text{Lower Control Limit} = \text{LCL} = \bar{p} - 3\sqrt{\frac{\bar{p}\bar{q}}{n}} = 0.093 - 3\sqrt{\frac{0.093 \times 0.907}{50}} = -0.03 \cong 0$$

$$\text{Control Limit} = \text{CL} = \bar{p} = 0.093$$

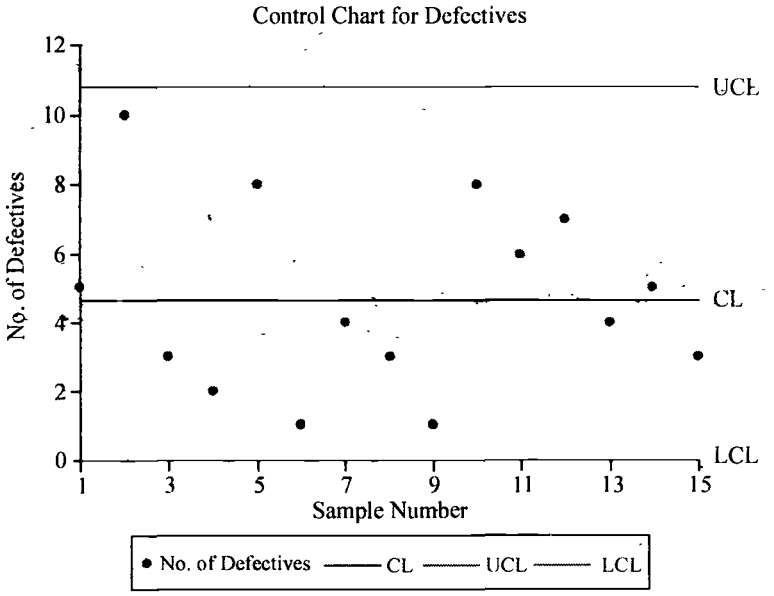


Fig. 78.1. *p chart for the data in Table 78.1*

The control chart shows that the number of defective in none of the samples is outside the control limits and accordingly we may comment that the system is within control.

78.3. Axes

X Axis: The axis is used to represent the sample numbers.

Y Axis: This axis is used to represent the values of the of fraction defectives in the sample.

78.4. Uses

(*a*) The plot is simple to draw and the calculations are relatively simple.

(*b*) The chart is used to discover the existence of any assignable cause of variation.

(*c*) The chart is successful even in case of small samples.

(*d*) The chart can deal with situations where the quality of the product cannot be measured but is only an attribute.

78.5. Related Techniques

(*a*) Process Control;

(*b*) Control Chart;

(*c*) σ Chart;

(*d*) $\bar{X}$ Chart; and

(*e*) *np* Chart.

79. Percent Defective Plot

79.1. Definition and Description

The plot is used to detect if, the percentage of defective items produced by a particular machine differs across the different sub samples. The plot can be used in industries by the quality control inspector to detect the performance of the machines. The plot is a line diagram between the sub-sample identification number and the percentage of defective in each sub-sample. In case of significant difference in the percentage of defectives in the sub-samples the points will be scattered, otherwise they will form an approximate straight line parallel to X axis. This can also be used to compare the performance of several machines producing the same product across various sub-samples. If multiple machines are compared, then in the graph we get a number of line diagrams equal to the number of machines to be compared. Also the performance of a machine before and after repair can be compared. In case of such a comparison we get two line diagrams in the graph.

79.2. Working Data

The working data is hypothetical and consists of the number of defective items in 12 sub-samples of size 50 each of a machine before and after overhaul. The data is also expressed in percentage.

Table 79.1: Number of defectives in different sub-samples by a machine before and after overhaul

Sub-sample No.	*Before Overhaul*		*After Overhaul*	
	No. of defective	*% of defective*	*No. of defective*	*% of defective*
1	0	0	1	2
2	4	8	1	2
3	1	2	1	2
4	2	4	2	4
5	1	2	4	8
6	3	6	7	14
7	1	2	7	14
8	3	6	2	4
9	2	4	2	4
10	2	4	6	12
11	3	6	1	2
12	4	8	2	4

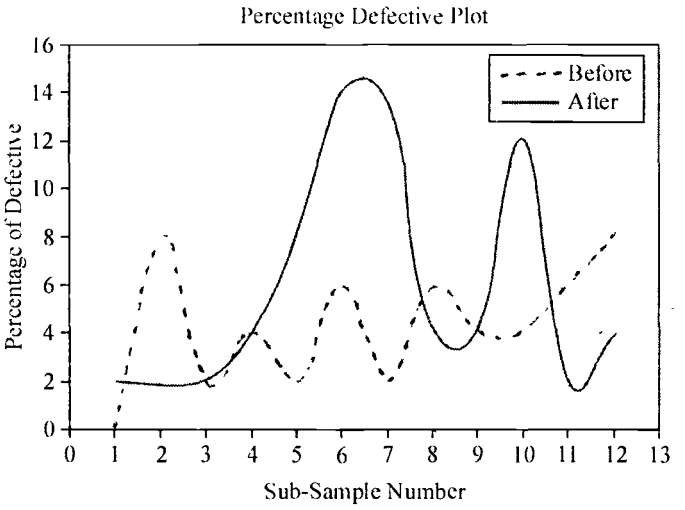


Fig. 79.1. *A Percentage defective plot for the data provided in Table 79.1*

Here we see that after the overhaul the number of defective has increased as well as the variation in the percentage of defective has also increased after the overhaul. Thus we may conclude that the overhaul has affected the machine adversely.

79.3. Axes

X Axis: The axis is used to represent the sub-sample identification number.

Y Axis: This axis is used to represent the percentage of defective.

79.4. Uses

(*a*) The plot provides a means to visualize the percentage of defective items in different sub-samples.

(*b*) If the sub-samples are collected over time then the plot can be used to understand the depression in the performance of the machine over the given time period.

(*c*) It can be used to compare the performance of the machine before and after overhaul.

(*d*) It helps in the comparison of the accuracy of production for several machines.

79.5. Related Techniques

(*a*) *p* Chart;

(*b*) Linear Intercept Plot; and

(**c**) Paired *t* Test.

80. Pictogram

80.1. Definition and Description

Pictograms are statistical diagrams in which the data are displayed with the help of pictures. They are very effective for propaganda purposes. They present dull masses of figures in interesting and attractive manner through objects of daily observation. The image of the entire data is fixed in the mind of the observer by a mere glance at the picture. Each symbol in the picture represents a definite numerical value. However the pictures selected for drawing the pictograms should be some way related to the subject matter of the display. Here the number of pictures representing a particular figure is proportional to the value. In some cases the pictures may even be fractional

80.2. Working Data

The data for drawing a pictogram is a hypothetical data supposed to have taken from a college where a considerable increase in the female enrollment over the years has been noticed. The figures are as follows:

Table 80.1: Yearwise female enrollment in a college

Year	2000-2001	2001-2002	2002-2003
Female Enrollment	230	350	475

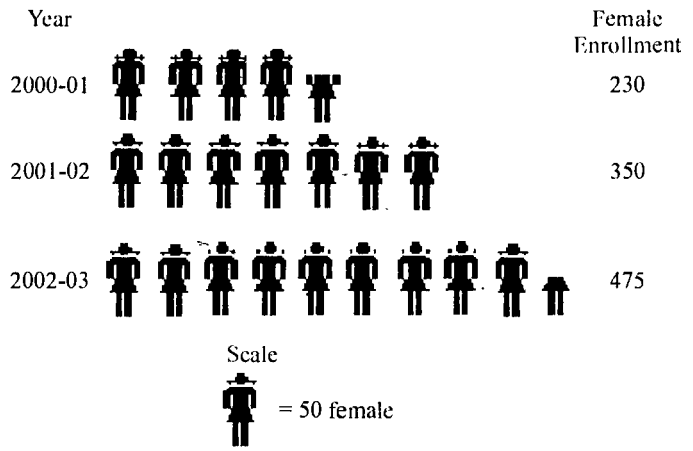


Fig. 80.1: *Pictogram for data in Table 80 1*

Here the picture of a girl is used for the purpose of representation. Each picture represents 50 females. Thus enrollment in 2000-01 is represented by 4 and a fraction, and so on for the other years.

80.3. Axes

In this plot no axes are used.

80.4. Advantages

(*a*) This plot does an excellent work of comparing the values through visualization.

(*b*) Pictograms are very attractive.

(*c*) The plot is very popular amongst people who find it difficult to understand numbers.

80.5. Disadvantages

(*a*) Special purpose software is required for producing such graphs. Commonly used statistical software does not provide this graph.

(*b*) Fractional figures may be difficult to produce manually.

(*c*) Can be used for representation of a single variable with few numbers of observations only.

(*d*) Time consuming display if produced manually.

80.6. Related Techniques

(*a*) Cartogram; and

(*b*) Pie-pictogram.

81. PIE DIAGRAM

81.1. Description of the Plot

Pie diagrams are two-dimensional diagrams. They are in the form of circles. A circle can be sub-divided into sectors and this can be easily done as the areas of the sectors are proportional to the angles drawn at the center. As there are 360 degrees at the center each degree of the angle represent $1/360^{th}$ area of the circle. Taking this fact into consideration, sectors can be drawn to represent the different component parts. A circle is drawn with convenient radius and different sectors are drawn corresponding to the various components with angles at the center. Thus the pie diagram consists of a circle which is sub-divided into several sectors by drawing radius. The area of the sectors is proportional to the values of the components. The different components are colored or shaded differently. The diagram will be more appealing if different colors are given to different sectors. The legend accompanying the graph can be used for indexing the various shades/colours.

81.2. Working Data

The working data gives the reserve of Petroleum (in %) for different regions. America 13, Europe 12, Africa 9, Middle East 61, Others 5.

81.3. Calculations and Figure

To draw a pie diagram we have to convert the figures to corresponding degrees, using the following formula :

If x is the percentage of petrolium under a particular head and if Σx is the total percentage then corresponding degrees are:

$$(x/\Sigma x) \times 360^{\circ}$$

Table 81.1: Calculations Related to Pie Diagram

Petrolium (%)	*Corrosponding*	*Degrees*
America	13	46.8
Europe	12	43.2
Africa	9	32.4
Middle East	61	219.6
Others	5	18
Total	100	360

81.4. Axes

This diagram does not require any axes for its representation. Thus the diagram may be drawn in plain paper and not in graph sheets.

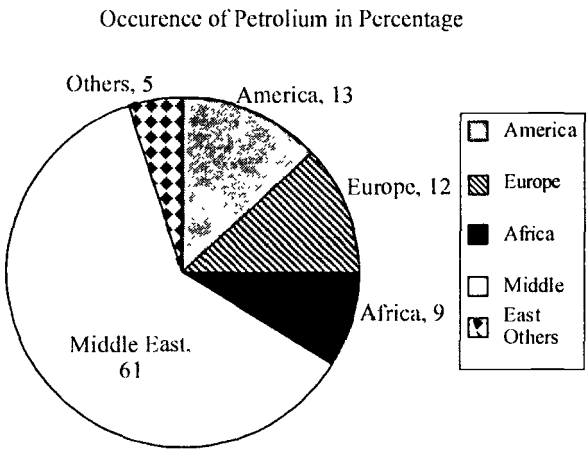


Fig. 81.1 *Pie Diagram based on data Table 81.1*

81.5. Uses

(*a*) This is probably the most commonly used diagram. Pie diagrams are commonly used in news papers, magazines etc.

(*b*) Such diagrams are widely used for comparing relative contribution of the different components under a particular head.

(*c*) They are simple to draw and the calculations involved are relatively simple, also, almost all the statistical software provides the option of drawing it.

81.6. Related Techniques

(*a*) Simple Bar Diagram;

(*b*) Sub-divided Bar Diagram;

(*c*) Histogram; and

(*d*) Impulse Chart.

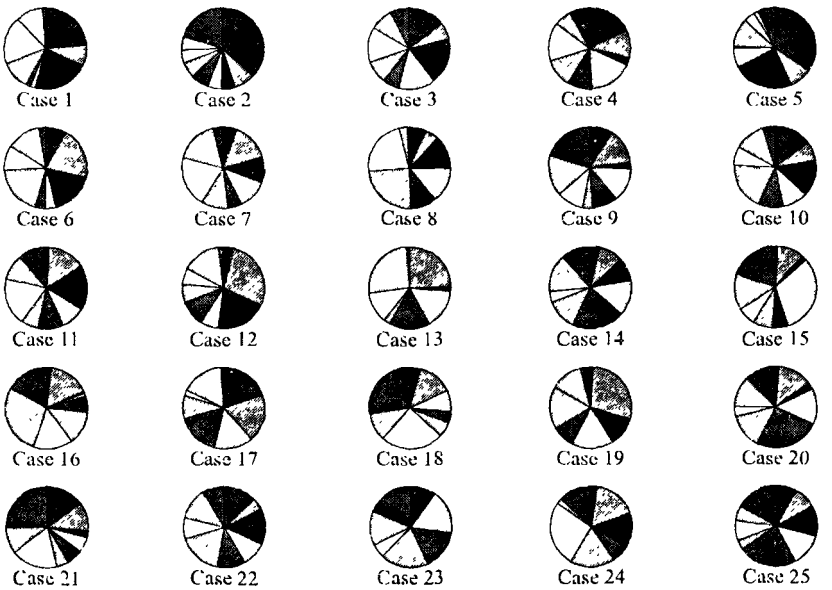
82. PIE ICON PLOT

82.1. Definition and Description

This is a plot used to display multivariate observations. In a pie icon plot each observation is represented by a pie diagram. In each pie, the circle is divided into several sectors by drawing lines from the centre to the radius such that area of each of the sector is proportional to the values of the variable it represents. Each of the sectors is differently colored so that the individual variables can be easily identified. Thus for n observations we will get the same number of pie diagram in the entire display. This display is used to understand the contribution of each of the variables in a particular observation. It should not be used for comparing the absolute values of a variable across the different observations but their proportions can be.

82.2. Working Data

The data used for this purpose is provided in Table 22.1. The data set is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups is given. However the data is standardized using the usual procedure.



Legend (clockwise): R_MEAT, W_MEAT, EGGS, MILK, FISH, CEREALS, ST_FOOD, PULSES, FRUITS.

Fig. 82.1. *The pie icon plot for the protein consumption data*

82.3. Axes

In this plot no axes are used so the order of the variables in a particular pie is considered either in the clockwise or in the anti-clockwise direction. In this case we have arranged them in clockwise fashion.

82.4. Advantages

(*a*) A pie icon plot is easier to interpret.

(*b*) In case of pie icon plot an additional variable will lead to an additional sector in all the pies and so it will not occupy any extra space unlike other icon plots like column icon plot or profile plot etc.

(*c*) This display is used to understand the contribution of each of the variables in a particular observation.

(*d*) STATISTICA has the option of drawing such a plot.

82.5. Disadvantages

(*a*) It should not be used for comparing the absolute values of a variable across the different observations. However one may standardize the data in order to perform the comparison.

(*b*) The plot is difficult to draw manually, one has to take the help of some statistical software.

82.6. Related Techniques

(*a*) Profile Icon Plot;

(*b*) Star Icon Plot;

(*c*) Chernoff Faces; and

(*d*) Column Icon Plot.

83. POISSONNESS PLOT

83.1. Description of the Plot

This plot is used to check if a given data set can be considered to have come from a Poisson distribution. The plot was introduced by Hoaglin (1980). We know that if X follows Poisson distribution with parameter λ then we have,

$$P(X=x) = \frac{e^{-\lambda}\lambda^{x}}{x!}, \lambda > 0, x = 0, 1, 2, \ldots$$

Now, if the total frequency is N, then the expected frequencies of the distribution corresponding to $x = 0, 1, 2, \ldots$ etc., are

$$\eta_x = N \,.\, P(X=x) = N \cdot \frac{e^{-\lambda}\lambda^{x}}{x!},$$

Taking the logarithm of both sides with respect to base 'e' we have,

$$\log \eta_x = \log N - \lambda + x \log \lambda - \log (x)!$$

$$\Rightarrow \quad \log \eta_x + \log (x)! = \log N - \lambda + x \log \lambda$$

$$\Rightarrow \quad \log [\eta_x (x)!] = \log N - \lambda + x \log \lambda \qquad \ldots(83.1)$$

Thus, from (83.1) it is clear that there is a linear relationship between x and $log(\eta_x (x)!)$. Now, if log $(\eta_x (x)!)$ is plotted against x then we get a straight line provided the data comes from a Poisson distribution. The fitted line has a slope of '$\log \lambda$' and an intercept of '$\log N - \lambda$'. The slope of the fitted line can thus be used as an estimate of 'λ'.

For drawing the poissonness plot, we compute the value of log $[o_x (x)!]$, where o_x denotes the values of the observed frequencies. The points $(x, \log [o_x (x)!])$ are then plotted in a graph paper along with the line (83.1). In case the observed data follows a Poisson distribution then, the plotted points *i.e.*, $(x, \log [o_x (x)!])$ lies in close proximity to the line (83.1). The value log $[\eta_x (x)!]$ is called as the count metameter and is denoted by $\phi (\eta_x)$.

83.2. Working Data

The data used for this purpose is taken from Student (1906) and is provided in the Table 83.1 given below:

Table 83.1: Count and corresponding frequencies from Student (1906)

Count:	0	1	2	3	4	5	6
Frequency:	103	143	98	42	8	4	2

83.3. Calculations

The corresponding table *i.e.*, Table 83.2 shows the relevant calculations required for a Poissonness plot and the Fig. 83.1 is the corresponding Poissonness plot to the data which shows that the data is a fairly good fit to the Poisson distribution.

Table 83.2: Calculation of expected frequencies, observed and expected count metameter values based on Student (1906) data

Counts (x)	*Observed Frequency* (o_x)	*Expected Frequency* (η_x)	$log\,[\eta_x(x)!]$	$log\,[o_x(x)!]$
0	103	106	2.025306	2.012837
1	143	141	2.149219	2.155336
2	98	93	2.269513	2.292256
3	42	41	2.390935	2.401401
4	8	14	2.526339	2.283301
5	4	4	2.681241	2.681241
6	2	1	2.857332	3.158362
	Σo_λ =400	$\Sigma\eta_\nu$ = 400		

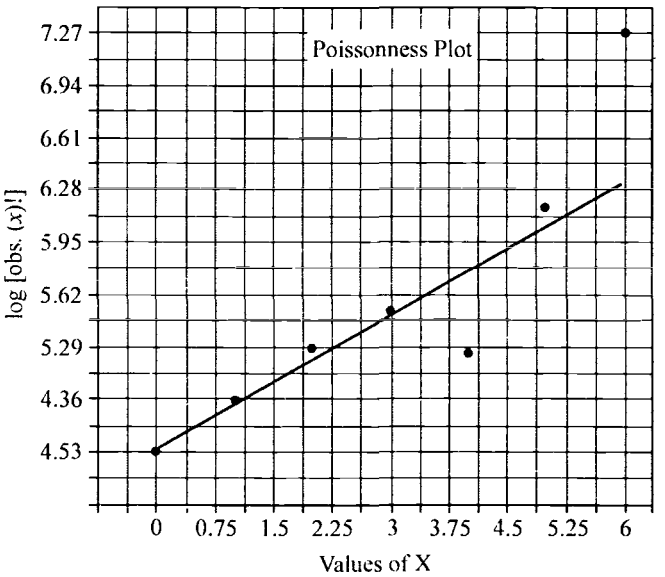


Fig. 83.1. *A Poissonness plot for the above data*

In this plot the dots represents the values of $log(o_x(x)!)$ *corresponding to the values of x. The straight line is* $log\,[\eta_x(x)!] = log\,N - \lambda + x\,log\,\lambda$. *The closer the red dots are to the line the better is the fit. However, the right tail of the straight line is very sensitive, small departure in the expected and observed counts at that end shows large departure in the graph. Since Poisson distribution deals with rare events so the value of* λ *will in general be small so the observed frequencies will be handsome for the lower values of x but it decreases considerably as the value of x increases. This makes the plot sensitive at the right tail.*

83.4. Axes

X Axis: It represents the value of the independent variable. In the figure we take the counts along the axis.

Y Axis: The vertical axis consists of the values of the count metameter function which is given by $\log\left[(x)!\eta_x\right]$.

83.5. Uses

(*a*) This plot as the name indicates can be used for checking the goodness of fit of the data to a Poisson distribution. The use is restricted and it is difficult to use the plot in deciding about the lack of fit by just looking at the Poissonness plot. Some confidence bands around the observed line can be a step towards the greater acceptability of the technique.

(*b*) With a little modification the plot can be used to detect the goodness of fit of a data to a truncated Poisson distribution.

(*c*) Probably none of the statistical software posses the plot by default.

83.6. Related Techniques

1. Binomialness Plot;
2. Chi-square test for Goodness of Fit; and
3. Probability Plot.

84. PROBABILITY PLOT

84.1. Definition and Description

Probability plot is a graphical technique for accessing whether sample data conform to a hypothesized distribution, based on subjective visual examination of the data. Our purpose is to test if the data follows a given theoretical distribution or not. The data is plotted against the theoretical distribution, in such a way that, the points should fall approximately in a straight line. If the plotted values show departure from the straight line then it implies that the data does not follow the specific distribution. The intercept and slope of the straight line are useful in the sense that they are the estimates of the location and scale parameters of the distribution. Though it is not so much important for the normal distribution, as the mean and standard deviation are the estimates of the location and the scale, it can be of great use in case of other distributions.

Let $y_1, y_2, \ldots, y_n$ be a random sample of size n. We are to test if the random sample comes from the null distribution $F_0(y)$. This is the probability distribution of Y given by the null hypothesis. The steps for the construction of a probability plot are as follows:

(*a*) Sample is first ordered or arranged in increasing order of magnitude. The arranged sample is thus denoted by $y_{(1)}, y_{(2)}, \ldots y_{(n)}$.

(*b*) The hypothetical probability $F_0(y_{(i)})$ for each $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ are calculated.

(*c*) We then apply continuity correction to the indicator variable i by subtracting $\frac{1}{2}$ from each of them.

(*d*) The values $(F_0(y_{(i)}), i - \frac{1}{2})$ are then plotted in a graph sheet. Alternately one can plot $(y_{(i)}, i/(n+1))$. If the plotted points fall approximately in a straight line then it may be concluded that the random sample comes from the null distribution $F_0(y)$, otherwise not.

84.2. Working Data

The data here is generated in MS Excel, is a set of random numbers $(y_i, i = 1, 2, \ldots, n)$ from a uniform distribution $(U(0, 1))$, which is then arranged to get the ordered sample $(y_{(i)}, i = 1, 2, \ldots, n)$. The values of $(F_0(y_{(i)}), i - \frac{1}{2})$ and the expected frequencies are then calculated and tabulated.

Table 84.1: Data and calculations for probability plot

y_i	$y_{(i)}$	$F_0(y_{(i)}) = y_{(i)}$	$i - (1/2)$	*Exp. freq*
0 449568	0.048341	0.048341	0 5	0.96682
0.764733	0.102756	0.102756	1.5	2.05512
0.137883	0.118686	0.118686	2.5	2.37372
0.272835	0.137883	0 137883	3 5	2.75766
0.973296	0.231117	0.231117	4.5	4.62234
0.450179	0.251228	0.251228	5 5	5.02456
0.102756	0.272835	0.272835	6.5	5.4567
0.941557	0.396435	0.396435	7.5	7.9287
0.396435	0.449568	0.449568	8.5	8.99136
0.535783	0.450179	0.450179	9.5	9.00358
0.853236	0.451827	0.451827	10.5	9.03654
0.451827	0.507218	0.507218	11.5	10.14436
0.231117	0.535783	0.535783	12.5	10.71566
0.983642	0.764733	0.764733	13.5	15.29466
0 507218	0.853236	0.853236	14 5	17.06472
0.936522	0.87701	0.87701	15.5	17.5402
0.118686	0 936522	0.936522	16.5	18.73044
0.251228	0.941557	0.941557	17.5	18.83114
0.048341	0.973296	0.973296	18.5	19.46592
0.87701	0.983642	0.983642	19.5	19.67284

Probability Plot Area A Random Data
From Uniform Distribution

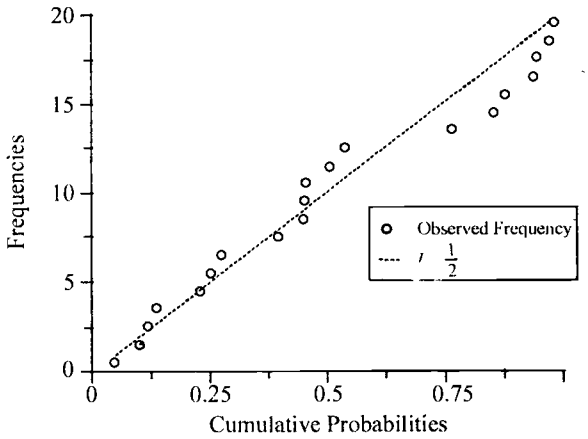


Fig. 84.1. *A probability plot for the data in Table 84.1*

From the plot we see that the observed frequencies approximately follow the straight line and so we may conclude that the set of random numbers shows a nearly good fit. We then perform one sampled K-S test, and find out if the data follows uniform distribution

84.3. Axes

X Axis: The axis is used to represent the cumulative probabilities of the observations.

Y Axis: This axis is used to represent the number of observations that are expected to lie below the corresponding cumulative probability.

84.4. Advantages

(*a*) Probability plots provide a quick check to the goodness of fit.

(*b*) Probability plots perform well both with large and small samples unlike other statistical tests which have strict restriction on sample sizes. Like the Chi-Square test for the goodness of fit test, which, require at least 50 to 100, observations for meaningful tests or the Wilks Shapiro test, which can be used only when the sample size lies between 10 to 1000.

(*c*) Probability plot can be used for a large variety of distributions.

(*d*) The intercept and slope of the straight line, which the plotted points approximately follow, can be used as the location and scale parameter of the distribution.

84.5. Disadvantage

The most striking problem with the plot is that different people reach to different conclusions regarding the goodness of fit test. Since, no proper confidence bands are developed around the straight line obtained from the plotting of expected frequency from the hypothetical distribution, so accessing the goodness of fit sometimes, become difficult even for experts.

84.6. Related Techniques

(*a*) Chi-square test for the goodness of fit;

(*b*) Anderson-Darling test;

(*c*) Wilks Shapiro test;

(*d*) EDF plot; and

(*e*) Normal Probability Plot.

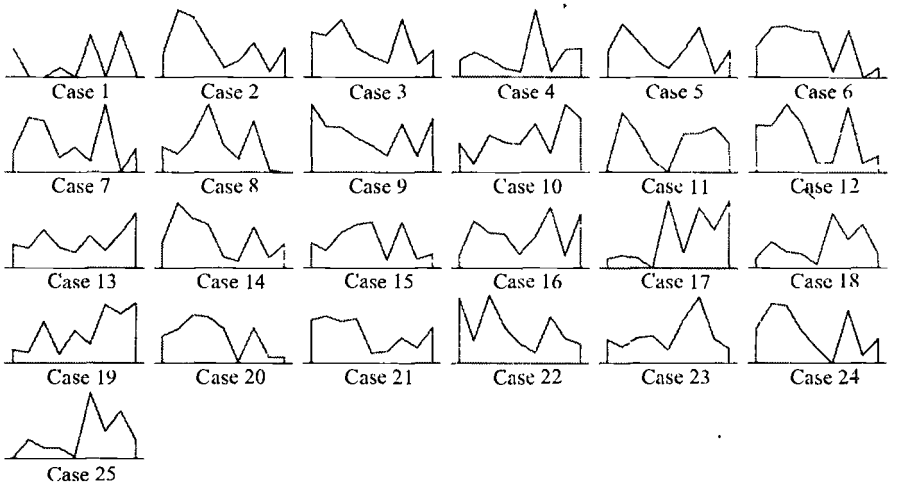
85. PROFILE ICON PLOT

85.1. Definition and Description

Profile icon plots are icon plots used to display multivariate observations. A profile icon plot is an extension of a histogram icon plot (57). In a histogram icon plot if the mid-points of the histogram are connected by straight line segments. The histograms are then removed from the plot and in such a case what remains is a polygon. The connected points are called as the vertices of the polygon. The plot thus formed is called as a polygon plot or a profile plot. Thus for a single observation the variable values will be represented by the height of the vertices of the polygon. A particular polygon will have p vertices one for each variable.

85.2. Working Data

The data used for this purpose is provided in Table 22.1. The data set is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups is given.



LEGEND (LEFT TO RIGHT): R_MEAT, W_MEAT, EGGS, MILK, FISH, CEREALS, ST_FOOD, PULSES.

Fig. 85.1. *The profile icon plot for the protein consumption data*

85.3. Axes

In this plot no axes are used so the order of the variables are considered either from left to right or from right to left.

85.4. Advantages

(*a*) A profile icon plot is easier to interpret.

(*b*) An important aspect of icon displays is partitioning of the observations into different clusters by visual detection only. This is easier to perform with profile icon plot compared to any other icon plot. At the initial stage one may start with dividing the polygons on the basis of skewness viz. positively skewed, negatively skewed, symmetrical and multi-model polygons.

(*c*) The display does not require the use of color.

(*d*) STATISTICA, S-plus, Stat Graphics have the option of drawing such a plot.

85.5. Disadvantages

(*a*) This plot is not much useful for the purpose of comparison of the performance of different variables across the observations.

(*b*) With increase in the number of variables the space required for display of an individual observation increases in case of a profile plot and hence it in not advantageous to represent more than five variables in those plots.

85.6. Related Techniques

(*a*) Star Icon Plot;

(*b*) Sunray Icon Plot;

(*c*) Chernoff Faces;

(*d*) Column Icon Plot;

(*e*) Pie Icon Plot; and

(*f*) Histogram Icon Plot.

86. Q-Q PLOT

86.1. Definition and Description

The quantile-quantile (Q-Q) plot is a graphical technique for determining if two data sets come from populations with a common distribution. A Q-Q plot is a plot of the quantiles of the first data set against the quantiles of the second data set. By a quantile, we mean the fraction (or percent) of points below the given value. That is, the 0.4 (or 40%) quantile is the point at which 40% percent of the data fall below and 60% fall above that value.

A 45-degree reference line is also plotted in the graph. If the two sets come from a population with the same distribution, the points should fall approximately along this reference line. The greater the departure from this reference line, greater is the evidence for the conclusion that the two data sets have come from populations with different distributions. The sample sizes of the two samples need not be equal.

86.2. Working Data

The sample data for the plot is taken from Anderson (1958). The data gives the measurement (in mm) of head length of the first two adult sons in a few families. The data is provided in Table 11.1

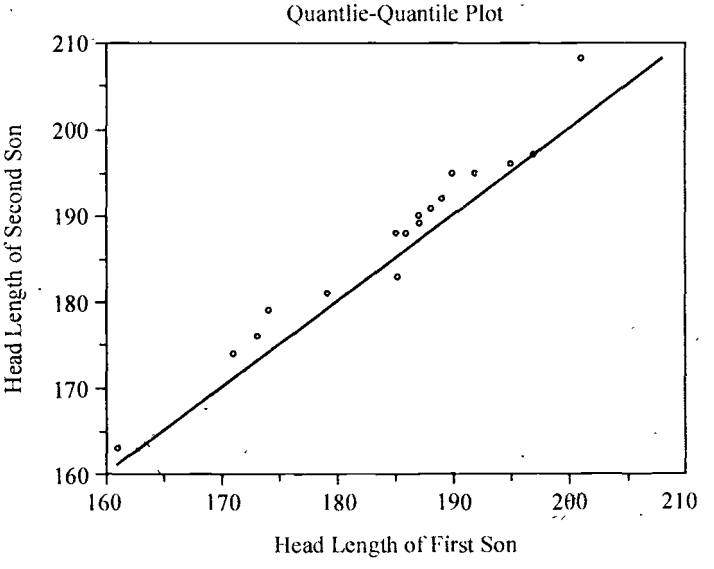


Fig. 86.1. ***A Q-Q plot for the data in Table 11.1***

The quantile-quantile plot shows that there is some discrepancy between the quantiles from the two data sets. However, in this case it is difficult to infer if the two data sets come from populations with a common distribution.

86.3. Axes

X Axis: The estimated values of quantiles from the first data set. Here head length of the first son is considered along the axis.

Y Axis: The estimated values of quantiles from the second data set. Here head length of the second son is considered along the axis.

86.4. Advantages

(*a*) Many distributional aspects can be simultaneously tested. For example, shifts in location, shifts in scale, changes in symmetry, and the presence of outliers can all be detected from this plot.

(*b*) The plot is used to check if the two data sets come from populations with a common distribution.

(*c*) The plot can be used to check if the two data sets have similar distributional shapes.

(*d*) It can also be used to compare the tail behaviour of two data sets.

86.5. Disadvantage

The most striking problem with the plot is that different people reach to different conclusions regarding the goodness of fit test. Since, no proper confidence bands are developed around the reference line so accessing the goodness of fit sometimes, becomes difficult even for experts.

86.6. Related Techniques

(*a*) Kolmogorov Smirnov two sample test;

(*b*) Chi-square test;

(*c*) Wilks Shapiro test;

(*d*) EDF plot; and

(*e*) Probability Plot.

87. R CHART

87.1. Definition and Description

R Chart is a type of control chart. Control charts are used to study the variability in the quality of a manufactured product and are used to understand if the process is within control. If the data is measurable and are usually continuous in nature then this type of control chart can be used.

Here k samples of the manufactured product are collected of same size 'n' (say) and the measurements are taken. Let x_{ij} be the value of the i^{th} ($i = 1, 2, ..., n$) observation in the j^{th} ($j = 1, 2, ..., k$) sample. Next, we compute the range (R_i) for each of the samples and accordingly the mean of the sample ranges *i.e.*, $\bar{R} = \frac{\sum R_i}{k}$.

A control chart consists of three lines parallel to the X axis, the control line, the upper control limit (UCL) and the lower control limit (LCL). Here we have,

Control Line = CL = $\bar{R}$

Upper Control Limit = UCL = $D_4 \bar{R}$

Lower Control Limit = LCL = $D_3 \bar{R}$

Where D_4 and D_3 are constants the values of which are obtained from table for corrosponding values of n. The tables are not provided in this book. Any standard book on applied statistics provides the same.

After all this calculations are done the values of R_i are plotted for different values of i. In other words the values of the ranges for the different samples are plotted in the form of dots against the sample numbers. The UCL and LCL are also drawn. If all the dots i.e. (i, R_i) fall within the UCL and LCL (control limits) then the system is said to be under control otherwise the system is beyond control.

87.2. Working Data

The data for this purpose is the life in hours of cells after complete charging obtained from a local manufacturer of cells. The data consists of 15 samples each of size 5.

Table 87.1: Life in hours of cells after full charging

Sample No.	*Life of Cells in Hours*					*Range* (R_i)
1	30.5	34 6	20.2	29.0	30.7	14.37
2	23.9	22.1	38.7	39 8	27.5	17.74
3	31.1	23:5	31.4	27.3	39.6	16.11
4	24.3	39.4	30.9	21.9	25.2	17.55

(contd...)

Sample No.	*Life of Cells in Hours*					*Range* (R_i)
5	20.7	20.6	37.7	26.3	32.3	17.14
6	36.6	27.1	20.1	33.8	29.9	16.46
7	37.7	27.9	20.2	25.7	36.2	17.50
8	26.7	38.4	22.4	29.4	24.1	16.04
9	29.7	20.3	30.6	35.2	36.1	15.83
10	22.8	33.3	32.3	33.0	27.2	10 49
11	25.9	22.4	29.3	28.2	21.6	7.73
12	30.1	31.6	24.4	27.1	27.7	7.24
13	22.1	26.6	38.3	33.2	29.6	16.24
14	38.8	26.6	27.6	24.5	37.5	14.28
15	39.0	36.2	35.9	33.7	22.0	17.03

Here, $\bar{R} = \frac{\sum R_i}{k} = \frac{221.76}{15} = 14.78$

Upper Control Limit = UCL = $D_4 \bar{R}$ = 2.11 × 14.78 = 31.19

Lower Control Limit = LCL = $D_3 \bar{R}$ = 0 × 14.78 = 0

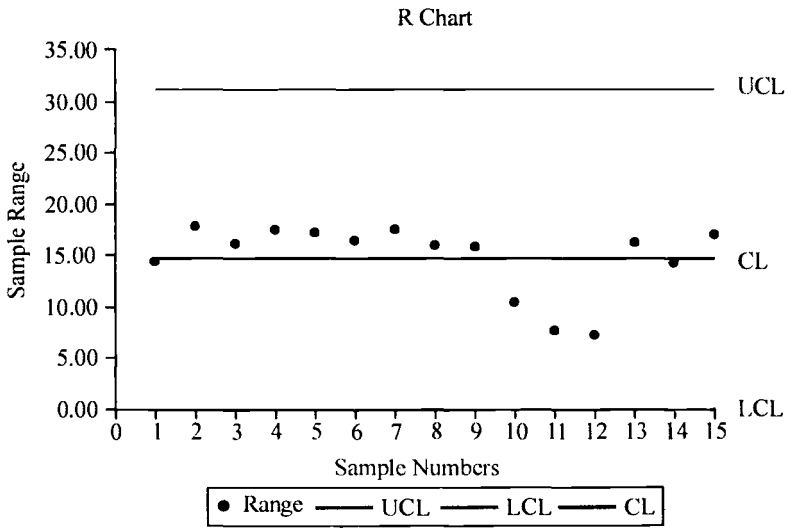


Fig. 87.1. *R chart for the data in Table 87.1*

The control chart shows that all the sample ranges are within control limits and accordingly we may comment that the system is within control.

87.3. Axes

X Axis: The axis is used to represent the sample numbers.

Y Axis: This axis is used to represent the values of the sample ranges.

87.4. Advantages

(*a*) The plot is simple to draw and the calculations are relatively simple.

(*b*) The chart is used to discover the existence of any assignable cause of variation.

(*c*) The chart is successful even in case of small samples.

87.5. Disadvantages

(*a*) The plot does not work well in case of larger samples.

(*b*) R chart is less sensitive than $\bar{x}$ chart.

(*c*) The chart cannot detect the extent of variability amongst the subgroups.

87.6. Related Techniques

(*a*) Process Control;

(*b*) Control Chart;

(*c*) σ Chart;

(*d*) $\bar{X}$ Chart;

(*e*) p Chart.

88. RADAR PLOT

88.1. Definition and Description

This is a plot in the form of a circle. A circle is drawn first and the values of the independent variable are considered as an angle corresponding to a pre determined radius of the circle. The response variable is considered along the radius. The plotted points are then added and accordingly a closed area across the radius is visible. Multiple numbers of response variables may also be plotted in the radar plot, producing several networks. In case of different response variables they may be differently colored for identification.

88.2. Working Data

The data provided in Table 78.1 is self explanatory and the plot below is the corresponding Radar plot to the data.

Table 88.1: Index of Industrial Production: Sector-wise

(Base Year : 1993 – 94 = 100)

Industry	*1997-98*	*1998-99*
Mining	126.4	125.4
Manufacturing	142.5	148.8
Food Products	133.8	134.7
Tobacco	158.1	178.5
Cotton textiles	125.6	115.9
Wool, Silk	172	176.8
Jute	114.3	106
Textile	158.7	153.1
Wood	128.5	121.0
Paper	146.4	169.8

Source: Economic Survey, 2000-01.

88.3. Axes

This plot does not require any axes for its drawing, only a radial scale is considered that can be demarked using concentric circles.

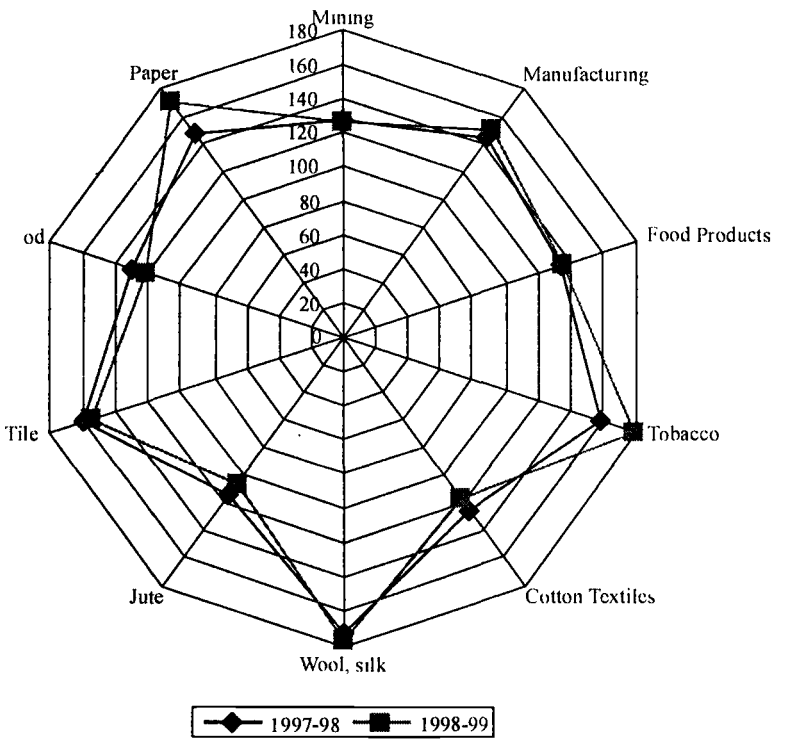


Fig. 88.1. *Radar Chart for the data in Table 88. 1*

88.4. Advantages

(*a*) The plot is easy to interpret.

(*b*) The plot occupies less space compared to an equivalent multiple bar diagram or multiple line diagram.

(*c*) The option of drawing this plot is available in MS Excel.

88.5. Disadvantages

(*a*) The plot is difficult to draw manually.

(*b*) Variation of the response variable corresponding to a particular category may be difficult to perform

(*c*) Large data sets are difficult to represent.

88.6. Related Techniques

(*a*) Line Diagram;

(*b*) Column Plot (Circular Base); and

(*c*) Star Icon Plot.

89. RESIDUAL HISTORGRAM

89.1. Definition and Description

Residual histogram is a graphical display technique used to differentiate between observed and fitted values through histograms. When histogram of a frequency distribution is drawn we can have a rough idea about the form of the probability distribution. If the plot is to check the normality of the data, and even if the histogram shows a bell shaped distribution, one cannot be sure of the normality of the data. This is because the normal curve is not the only bell shaped curve. To evaluate the distribution exactly, a comparison must come up between the observed frequencies and expected frequencies fit by the normal distribution.

Let n_i be the observed frequency and $\hat{n}_i$ be the corresponding expected frequency. Now $n_i - \hat{n}_i$ may be considered as the difference between observed and expected frequencies. So when the values of $n_i - \hat{n}_i$ are plotted against the classes intervals, in case of a reasonable good fit one would expect small difference of residuals in both sides of the x axis.

89.2. Working Data

For plotting the above-mentioned plots we take the data set from Gupta (1952) which is provided in Table 24.1. The data set showing the lifetime (in hours) of 300 electric lamps which follows normal distribution.

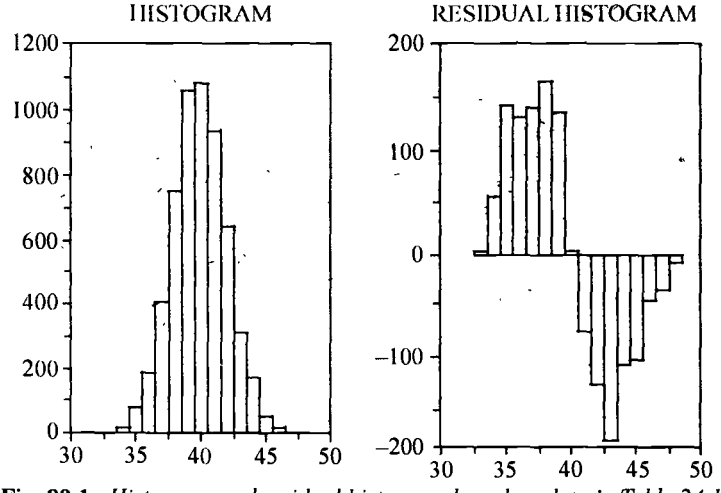


Fig. 89.1. *Histogram and residual histogram based on data in Table 24.1*

The expected frequencies are computed in the usual fashion and accordingly the values $n_i - \hat{n}_i$ are computed for each class interval. From the histogram it appears that the data follows a normal distribution but from the residual plot wide difference in observed and expected frequencies are noticed. Also a pattern is clear amongst the residuals belonging to the first and second half.

89.3 Axes

X Axis: It is used to represent the class intervals. For this data set Life Time (in hrs.) is taken along the X-axis.

Y Axis: The vertical axis is used for representing the residuals of frequencies along the different classes, *i.e.*, the values of $n_i - \hat{n}_i$ are taken.

89.4. Advantages

(*a*) They are very simple technique to judge the goodness of fit of a data set.

(*b*) The calculations involved are relatively simple. The plot can be easily understood and easily interpreted.

89.5. Disadvantages

(*a*) It is difficult to infer in all cases about the goodness of fit.

(*b*) The plot cannot be drawn for open-end classes.

But since the class frequencies are less in the tails so large differences between the observed and expected are noticed in the tails in case $n_i - \hat{n}_i$ are plotted against the class intervals.

89.6. Related Techniques

(*a*) Histogram;

(*b*) Residual Rootogram; and

(*c*) Chigram.

90. RESIDUAL PLOT

90.1. Definition and Description

Whenever we express a variable with that of the other in terms of a mathematical relation then the most general form in Statistics is given by

$$y = f(x) + e$$

Where, the variable 'e' is called as the error component associated with the model. Generally, we will not know the values of 'e' for all the given values of 'x'. Thus, we can estimate the coefficients involved in the function $f(x)$ and accordingly estimate the values of 'y' for a given 'x', which is denoted by '$\hat{y}$'. The difference 'y' and '$\hat{y}$' *i.e.*, '$y - \hat{y}$' is called as the residual. In case of a good fit these residuals are randomly distributed. The residual plot is a plot between the independent variable i.e. x and the corresponding residual *i.e.*, e.

90.2. Working Data

The data for this purpose is provided in Table 37.1, which comprises of some hypothetical values of X and Y. A linear function is fitted and accordingly the residuals are computed (Ref: Table 37.1).

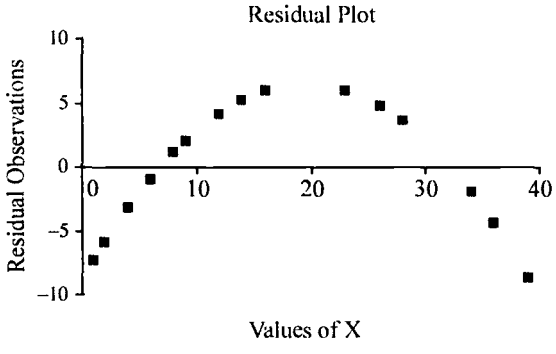


Fig. 90.1. *A Residual Plot for data in Table 37.1*

The residual plot shows that there is a pattern in the residuals and hence it cannot be considered as random observations. This gives us an indication that the model fitted is not a proper one for the data.

90.3. Axes

X Axis: The axis is used to represent the values of the independent variable.

Y Axis: This axis is used to represent the residuals from the fitted models.

90.4. Advantages

(*a*) The plot provides a quick check for the randomness of the residuals from a model.

(*b*) The randomness of the residuals can be studied both for large as well as small samples.

(*c*) In case of non-randomness in the residuals the plot suggests a better model for the data.

90.5. Disadvantage

The most striking problem with the plot is that different people reach to different conclusions regarding the randomness of the residuals. In some cases the non-randomness may be very easy to identify, but it is not always so easy for the human eye to guess. So accessing the randomness of residuals of the fit may sometimes be difficult even for experts.

90.6. Related Techniques

(*a*) Residual Histogram;

(*b*) Scatter Diagram;

(*c*) Detrended Probability Plot; and

(*d*) Serial Correlation Plot.

91. ROOTOGRAM

91.1. Definition and Description

Rootogram is a graphical display technique used to differentiate between observed and fitted values through histograms. When histogram of a frequency distribution is drawn we can have a rough idea about the form of the probability distribution. If the plot is to check the normality of the data, and even if the histogram shows a bell shaped distribution, one cannot be sure of the normality of the data. This is because the normal curve is not the only bell shaped curve. To evaluate the distribution exactly, a comparison must come up between the observed frequencies and expected frequencies fit by the normal distribution. Let n_i be the observed frequency and $\hat{n}_i$ be the corresponding expected frequency. Now $n_i - \hat{n}_i$ may considered as the difference between observed and expected frequencies. But since the class frequencies are less in the tails so large differences between the observed and expected are noticed in the tails in case $n_i - \hat{n}_i$ are used. So, Rice (1995) used a variance stabilizing transformation and proposed that $\sqrt{n_i} - \sqrt{\hat{n}_i}$ must be plotted against each class instead of $n_i - \hat{n}_i$. The plot thus obtained is also called as the residual rootogram.

91.2. Working Data

For plotting the above-mentioned plots we take the data set from Gupta (1952) which is provided in Table 24.1. The data set showing the lifetime (in hours) of 300 electric lamps which follows normal distribution.

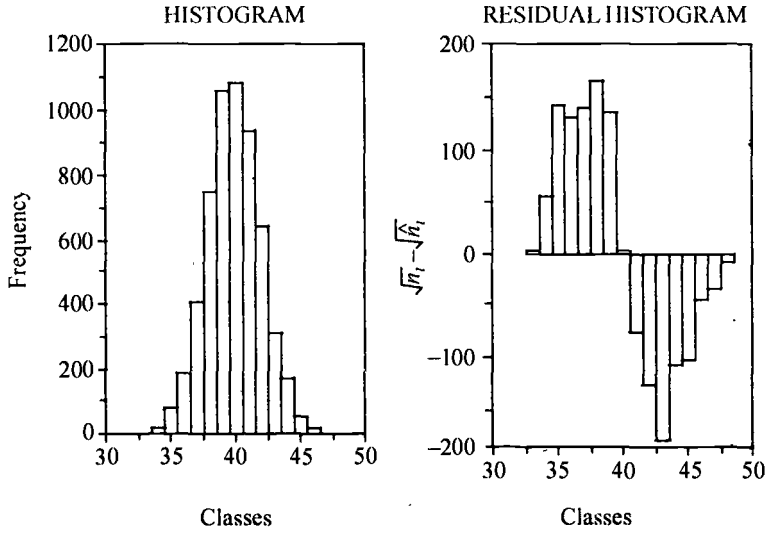


Fig. 91.1. *Histogram and residual rootogram based on data in Table 24.1*

The expected frequencies are computed in the usual fashion and accordingly $\sqrt{n_i} - \sqrt{\hat{n}_i}$ *values are computed for each class interval. From the histogram it appears that the data follows a normal distribution but in the residual rootogram the variance amongst the residuals are stabilized. The residuals are not randomly distributed and so a pattern in the residuals implies that the data is not a good fit to the distribution.*

91.3 Axes

X Axis: It is used to represent the class intervals. For this data set Life Time (in hrs.) is taken along the X-axis.

Y Axis: The vertical axis is used for representing the difference between root of frequencies, *i.e* , observed minus expected.

91.4. Advantages

(*a*) They are very simple technique to judge the goodness of fit of a data set.

(*b*) The calculations involved are relatively simple. The plot can be easily understood and easily interpreted.

91.5. Disadvantages

(*a*) It is difficult to infer in all cases about the goodness of fit.

(*b*) The plot cannot be drawn for open-end classes.

91.6. Related Techniques

(*a*) Histogram;

(*b*) Chigram; and

(*c*) Residual Histogram.

92. RUN SEQUENCE PLOT

92.1. Definition and Description

A very simple graphical tool which produces a summary of a univariate data set related to location, spread as well as the presence of outliers is the run sequence plot. Let Y be a variable taking the values $Y_1, Y_2,\ldots,Y_n$. Then the plot is produced by plotting the values of Y_i against the index variable i. The points are plotted and are not joined. Thus we get n points of the data set separated by an equal horizontal alignment.

92.2. Working Data

The data is taken from Jarrett (1979) which gives the time intervals between successive coal mining disasters killing atleast 10 people from 15th March 1851 to 22nd March 1962. The actual data consists of 191 values but the table shows the first thirty values only.

Table 92.1: Data related to Coal Mining Disasters

157	123	2	124	12	4	10	216	80	12
33	66	232	826	40	12	29	190	97	65
186	23	92	197	431	16	154	95	25	19

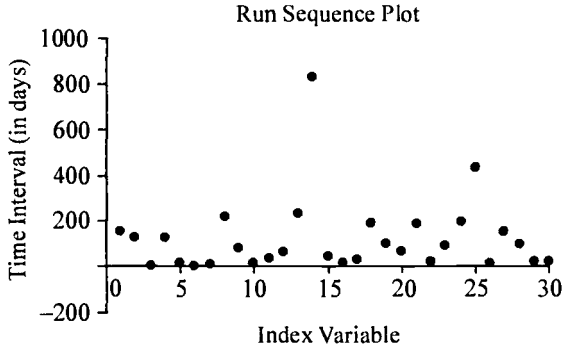


Fig. 92.1. *A Run Sequence plot for data in Table 92.1*

The plot shows that the disasters are frequent with Y_{14} being an outlier. The location value remains almost uniform throughout the plotted values.

92.3. Axes

X Axis: The index variables are taken along the X axis *i.e.*, $i = 1, 2, 3,\ldots$.

Y Axis: The values of the response variable are considered along the axis. Here time interval (in days) are considered along the Y axis.

92.4. Uses

(*a*) The plot helps us to understand if there is any shift in the location of the data.

(*b*) The plot also helps us to understand if there is any shift in the spread of the data.

(*c*) The presence of outliers can also be detected.

92.5. Related Techniques

(*a*) Scatter Diagram;

(*b*) Lag Plot; and

(*c*) Auto-Correlation Plot.

93. SCATTER DIAGRAM

93.1. Definition and Description

The scatter diagram is the simplest method of studying relationship between two variables, X and Y (say). The simplest device for ascertaining whether variables are related is to prepare a dot chart, where the horizontal axis represents one variable (X), and vertical axis representing the other (Y). Here we have the observations in pairs *i.e.*, of the form (x_1, y_1), $(x_2, y_2), \ldots, (x_n, y_n)$ where in each pair, the first value corresponds to X series and the other correspond to Y series. Each pair is then plotted in the graph in the form of a dot. The diagram so obtained is known as scatter diagram or dot diagram. From the scatter diagram one can have a fairly good idea about the relationship between variables. Scatter diagrams are two-dimensional diagrams. They are simple to draw and easy to interpret.

93.2. Interpretation of the Plot

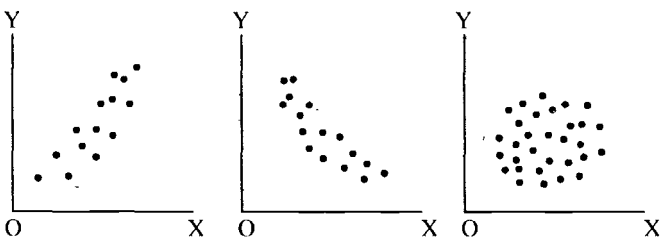
The following points may be borne in mind in interpretation of a scatter diagram:

(*a*) If the points are very close to each other, a fair amount of correlation may be expected between the two variables. On the other hand, if points are widely scattered, a poor correlation is expected between them.

(*b*) If the points on the scatter diagram reveal any trend (upward or downward), the variables are supposed to be correlated.

(*c*) If there is an upward trend rising from lower left hand corner to upper right hand corner, the correlation is positive. On the other hand, if there is a downward trend from upper left hand corner to lower right hand corner, the correlation is negative.

(*d*) If all the points lie on a straight line starting from left bottom and going upwards to the right top, then the correlation is perfectly positive. On the other hand, if all the points lie on a straight line starting from left top and going downwards to the right bottom, the correlation is perfectly negative.



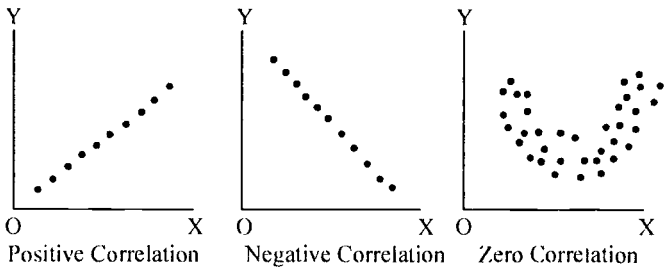


Fig. 93.1. *The scatter diagram for various types of correlation*

93.3. Working Data

The working data is taken from Altman (1991). The data set gives the values of a fat content (in percentage) in the body of 18 individuals along with their ages.

Table 93.1: Percentage of fat in human body and corresponding age

Age	23	23	27	27	39	41	45	49	50
% of fat	9.5	27.9	7.8	17.8	31.4	25.9	27.4	25.2	31.1
Age	53	53	54	56	57	58	58	60	61
% of fat	34.7	42	29.1	32.5	30.3	33	33.8	41.1	34.5

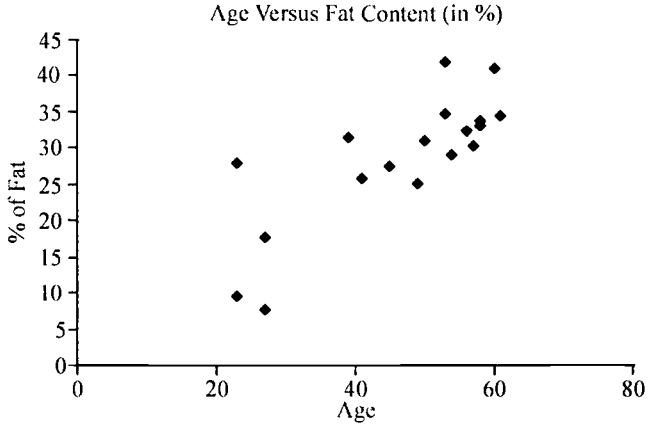


Fig. 93.2. *Scatter Diagram based on data Table 93.1*

From the scatter diagram we may conclude that there is a positive relation between the age of the individuals and the percentage of fat content, i.e., with increase in age fat content in human body is likely to increase.

93.4. Axes

X Axis: It represents the value of the independent variable. In the figure Age is independent and hence it is considered along the axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable or the dependent variable. For this case we took the fat content (in %) along the Y axis.

93.5. Uses

(a) This is probably the most commonly used diagram to understand the relationship between two variables.

(b) The plot can be used as diagnostic tool for understanding the mathematical relation *e g.* linear, quadratic, exponential, logarithmic etc. between the variables. Thus the user may fit the more appropriate curves instead of fitting all possible curves.

(c) Though not exact but this plot can be used for quickly deciding about the type of correlation between variables.

(d) Almost all the statistical software provides the option of drawing it.

93.6. Related Techniques

(*a*) Correlation Coefficient;

(*b*) Simple Regression; and

(*c*) Line Diagram.

94. SCATTERPLOT MATRIX

94.1. Definition and Description

A scatterplot matrix for p variables are obtained by arranging the scatterplots for all $p \times (p-1)$ ordered pairs of variables in a matrix. Each pair of variables X_i and X_j (say) are plotted twice once with X_i along the horizontal axis and X_j along the vertical axis and the other one in the reverse order with X_j along the horizontal axis and X_i along the vertical axis. The $p \times (p-1)$ scatter plots are then arranged like that of a matrix with p rows and p columns. Each plot is called as a panel. The panel belonging to the i^{th} row and j^{th} column is the scatter plot with X_i along the vertical axis and X_j along the horizontal axis. The diagonal panels are kept empty and in those panels the names of the variables are written, *i.e.*, name of first variable in the panel (1, 1), second variable in the panel (2, 2), and so on. However some software provide the option of drawing histograms in the diagonal panels. If there are p variables then there will be p diagonal panels and accordingly each diagonal panel consists of a histogram for a particular variable, giving an idea about the distribution of that variable. In the display the scale of the axes, tic marks associated with the axes etc. are concealed to give a clear view. Though the pioneer for the scatterplot matrix is not known but Tukey and Tukey (1981) dealt with an organized collection of scatter plots and they termed them as *draughtman's display.*

94.2. Working Data

The data used for this purpose is provided in Table 22.1. The data set is used for drawing the display is only an abridged form of that table. Only 5 food groups are considered for the drawing of the scatter plot matrix. The actual table is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups is given.

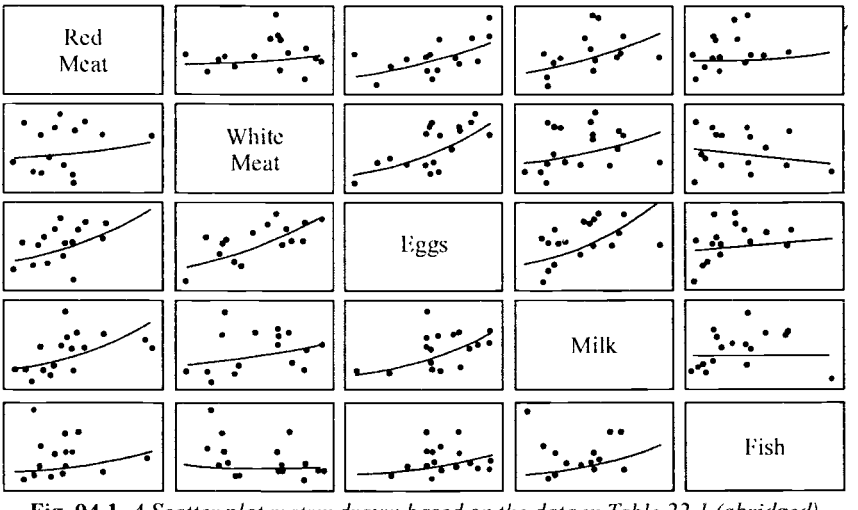


Fig. 94.1. *A Scatter plot matrix drawn based on the data in Table 22.1 (abridged)*

The scatter plot shows the fitted exponential curve to the data set which gives us an idea about the relation between the variables on an average.

94.3. Axes

This plot comprises of a large number of plots and each of these plots has its own axes.

94.4. Advantages

(*a*) This plot helps to study the relationship of each variable with the set of all the other variables by treating a particular variable as an independent variable as well as a dependent variable.

(*b*) If histograms are drawn in the diagonal panels then it helps in the multiple comparison of skewness, kurtosis and distributional pattern of a number of variables.

(*c*) Commonly available statistical software like SPSS, Statistica, Stata, S-plus etc. supports this plot.

94.5. Disadvantages

(*a*) Since in the display the scale of the axes, tic marks associated with the axes etc. are concealed so it sometimes becomes difficult to read the values from the figure and also the data can be easily manipulated.

(*b*) As the number variables increases the area available for each panel decreases so only a few variables can be considered for drawing this plot.

94.6. Related Techniques

(*a*) Scatter Diagram; and

(*b*) Trellis Display.

95. SIEVE'S DIAGRAM

94.1. Definition and Description

Sieve's diagrams are used for graphical representation of the two way contingency tables. Sieve's diagram is an extension over mosaic display originally developed by Hartigan and Kleiner (1981). Readers are requested to read mosaic plot for better understanding of Sieve's diagram. Riedwyl and Schiipbach (1983) proposed a mosaic display in such a way that the area of the tiles in the mosaic is proportional to the expected frequencies. This diagram was later called as the 'Parquet Diagram'. In this display the height of each rectangle is proportional to its row total (r_i) and width of each rectangle is proportional to its column totals (c_j). Thus, the area is proportional to $r_i \times c_j$ which also is proportional to the expected frequency. Thus, we see that rectangles representing the values in each row have the same height. Similarly the rectangles representing values in each column has the same width and thus remains aligned horizontally.

95.2. Working Data

The working data is provided in Table 69.1.The data is a two way contingency table derived from a 5-way contingency table used in Edwards and Kreiner (1983).

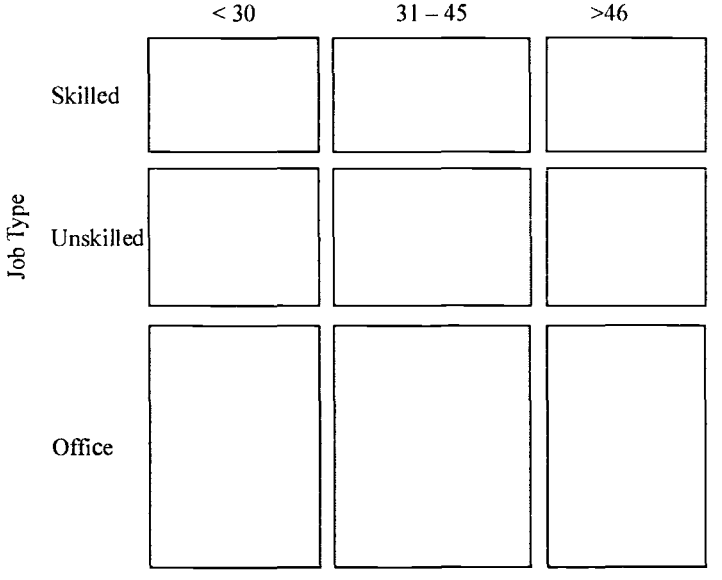


Fig. 95.1. *A Sieve diagram for data in Table 69.1*

Here the area of each rectangle is proportional to the expected cell frequencies.

95.3. Some Additional Insight

The basic plot does not have any scheme to display the difference between the observed and expected frequencies. Riedwyl and Schiipbach (1994) attached this by a modification to the plot. The modification was to divide each tile into small squares. The squares in each rectangle are equal to the observed frequencies. This helps, one to understand the difference between the observed and expected frequency, which appears from the density of shading. Also colors were used to indicate whether the deviation of the observed frequency (o_{ij}) from the expected (e_{ij}) are positive or negative. For example, negative residuals may be colored red and positive residuals blue. Thus, this extended display visualizes the observed frequency, expected frequency and the difference between the two.

95.4. Advantages

(*a*) The plot is used for the visualization of expected frequencies of various cells in case of data arranged in the form of contingency table.

(*b*) The plot is simple to draw and its interpretation is also easy.

95.5. Disadvantages

(*a*) The display does not support visualization of the observed cell frequencies.

(*b*) This display does not support any comparison between the observed and expected frequencies.

(*c*) Commonly used statistical software does not provide the option to draw this plot.

95.6. Related Techniques

(*a*) Chi-square test for independence in contingency table;

(*b*) Association Plot;

(*c*) Mosaic Plot; and

(*d*) Four Fold Display.

96. STACKED LINE CHART

96.1. Definition and Description

This is a multiple line chart. But here each value of a particular response variable is obtained by adding with the previous one, in other words the cumulative totals of the different response variables are obtained. The values of the first response variables are plotted and then the plotted points are joined by straight lines. The values of the second response variable are then added to the first and the plotting is done. This process is repeated for each of the response variable. However it should be sensible to add together the data sets in order to form the cumulative totals. Obviously the data should be of same unit.

96.2. Working Data

The data used for this plot is taken from Economic Survey of India, 1999-2001. The table shows the actual stock of food grains a hypothetical one representing data related to the production of an industry for different years.

Table 96.1: Actual stock of wheat in million tons

1	*1999* *2*	*2000* *3*	*2001* *4*	*Cumu. Tot.* *(2) + (3)*	*Cumu. Tot.* *(2) + (3) + (4)*
January	12.7	17.2	29.9	25.0	54.9
April	9.7	13.2	22.9	21.5	44.4
July	22.5	27.8	50.3	38.9	89.2
October	20.3	26.9	47.2	36.8	84

96.3. Axes

X Axis: The axis is used to represent the independent variable. Here different quarters are used.

Y Axis: This axis is used to represent the values of the response variable. The variable is Stock of Wheat in this case.

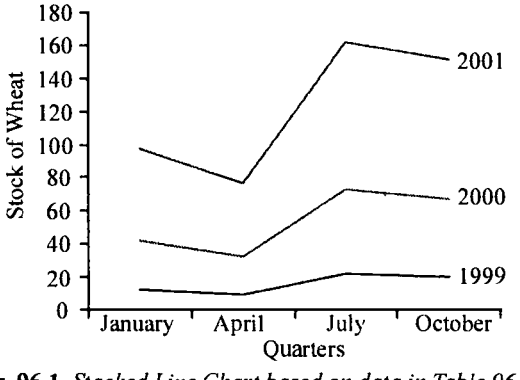


Fig. 96.1. *Stacked Line Chart based on data in Table 96.1.*

96.4. Advantages

(*a*) The plot provides a visual check related to the consistent variation in the data set.

(*b*) Here the lines do not cross each other and so each response variable can be studied easily.

(*c*) The graph can be drawn even without the use of color.

96.5. Disadvantages

(*a*) The plot cannot be used for the purpose of comparison of different points of the graph as there is no stable base line.

(*b*) The central values of the data set cannot be compared from this plot.

96.6. Related Techniques

(*a*) Line Diagram;

(*b*) Area Chart;

(*c*) Multiple Line Diagram; and

(*d*) Historigram.

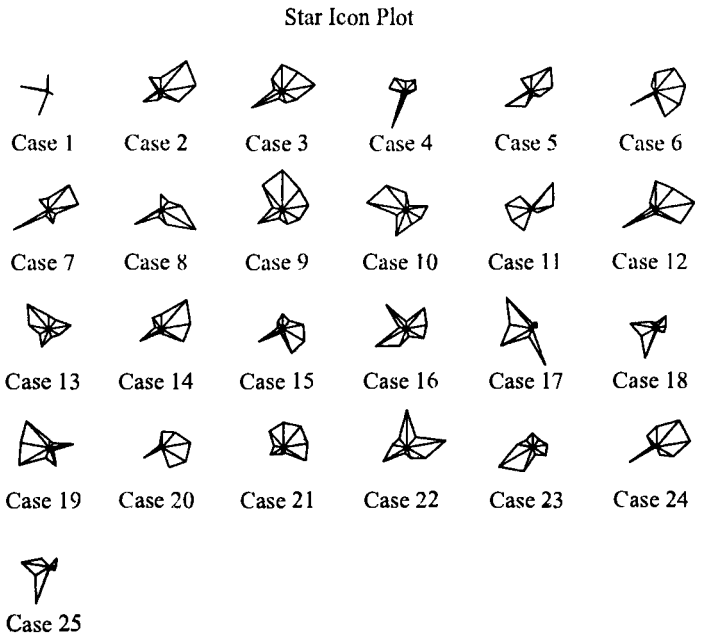
97. STAR ICON PLOT

97.1. Definition and Description

Star icon plots are icon plots used to display multivariate observations. Here corresponding to each multivariate observation we draw a central point with some rays coming out from it. Here, each variable in the observation is represented by a ray that comes out form a central point. These rays are equally spaced around the central point. The angle between any two adjacent rays would be $360^0/p$, where p is the number of variable in the system. The length of the rays is proportional to the value of the variable it represents. The heads of the rays are then connected by line segments to complete the star. According to Henry (1995) - "Stars are not literally stars, but plots that end up as polygons having as many sides as the number of variables to be plotted, with three variable plotted as triangles as the minimum."

97.2. Working Data

The data used for this purpose is provided in Table 22.1. The data set is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups is given.



LEGENG (CLOCKWISE): R_MEAT, W_MEAT, EGGS, MILK, FISH, CEREALS, ST_FOOD, PULSES, FRUITS,

Fig. 97.1. *The star icon plot for the protein consumption data*

97.3. Axes

In this plot no axes are used so the order of the variables in a particular star is considered either in the clockwise or in the anti-clockwise direction. In this case we have arranged them in clockwise fashion.

97.4. Advantages

(*a*) A star icon plot is easier to interpret. If the size of the star is large then we can understand that the observation has high values of the variable and vise versa. A symmetrical star implies that the observation performs almost equally in case of all the variables.

(*b*) In case of star icon plot an additional variable will lead to an additional ray that will come out from the central point and will not occupy any extra space unlike other icon plots like column icon plot or profile plot etc.

(*c*) STATISTICA has the option of drawing such a plot.

97.5. Disadvantages

(*a*) Since in a star plot the rays correspond to different variables so each ray will have a different scale of measurement. Thus we cannot compare the performance of the different variables within an observation. To get rid of this problem one may think of standardizing the variables and draw the plot.

(*b*) The variables in a particular observation often become difficult to identify separately if all of the rays are of the same color. This can be avoided in case separate color is used for each variable.

97.6. Related Techniques

(*a*) Profile Icon Plot;

(*b*) Sunray Icon Plot;

(*c*) Chernoff Faces;

(*d*) Column Icon Plot; and

(*e*) Pie Icon Plot.

98. STEM AND LEAF DIAGRAM

98.1. Definition and Description

A stem and leaf plot, or stem plot, is a technique used to represent either discrete or continuous variables in a semi graphical display. A stem and leaf plot is used to organize data at the time they are collected and reduce the time of noting down all the numerical values. A stem and leaf plot looks something like a bar graph extended along the horizontal axis. Each number in the data is broken down into a stem and a leaf, thus the name. The stem of the number includes all except the last digit. The leaf of the number will always be a single digit. In case of numbers after decimal points it may be rounded of if the user thinks so. For drawing the diagram for a set of numerical values on the left hand side of the page, write down the thousands, hundreds or tens (all digits except the last one). These are called as the stems. A line is drawn to the right of these stems. On the other side of the line, write down the last digit of a number. These are called as the leaves.

98.2. Working Data

For drawing the diagram let us consider the marks of 20 students taken from a class obtained from a total of 500. The marks are 356, 370, 419, 422, 333, 207, 350, 355, 372, 375, 378, 335, 337, 338, 411, 427, 293, 355, 297 and 295.

Fig. 98.1. *A Stem and Leaf diagram to the data provided above*

Stem	*Leaf*
20	7
29	3 7 5
33	3 5 7 8
35	6 0 5 5
37	0 2 5 8
41	9 1
42	2 7

The leaves may be ordered in some cases i.e arranged in ascending or descending order of magnitude.

98.3. Axes

The display does not require the use of any axes.

98.4. Uses

(*a*) It acts as a quicker way of noting down the values of the observation.

(*b*) It is better than a frequency table as it does not conceal the actual value of the observations.

(*c*) It can be used to understand the spread of data and is also useful in understanding the highest, lowest and the outliers.

98.5. Related Techniques

(*a*) Histogram; and

(*b*) Simple Bar Diagram.

99. SUNFLOWER PLOT

99.1. Definition and Description

The sunflower plot can be considered as an extension of a glyph plot. The plot can be used to represent three variables two of which are numerical variable and the third one is either a numerical variable or a categorical variable. At the very out set one independent variable and other response variable is plotted like that of a scatter plot. The plotting symbol is generally 'o'. The third variable is represented by whiskers (or rays) that come out from each of the plotted symbols. The number of whiskers that comes out from each symbol is proportional to the value of the third variable. The higher the value of the third variable more is the number of whiskers (or rays) that the plotted point carries. The length of the whiskers is identical and spreads in all directions producing the effect of a flower and hence the name sunflower plot. In case the third variable is a categorical variable then it may be represented by one ray from the plotting symbol for the first category, two rays from the plotting symbols for the second category and so on.

Cleveland and McGill (1984) introduced the plot. Beherns and Yu (1995) discuss such a plot and also about the software packages in which the sunflower plot is available. Dupont and Plummer (2002) extended the plot even further by drawing either a light sunflower or a dark sunflower -where in the light sunflower one petal represents one observation and in the dark sunflower each petal represents k observations.

99.2. Working Data

The data used for drawing the plot is same as that used to draw the bubble plot that is given in Table 17.1. The data comprises of heights, weights and age of 30 individuals collected non-randomly from some known individuals.

99.3. Axes

X Axis: It can represent the values of the variable that we suspect may have a relation to the response variable. Here the height is represented along the X axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. Here weight is represented along Y axis.

The third variable is represented by whiskers (or rays) that come out from each of the plotted symbols. The number of whiskers that comes out from each symbol is proportional to the value of the third variable, *i.e.*, age in this case. A maximum of eight petals are allowed to come out from a sunflower. The more the age more is the number of petals from each sunflower.

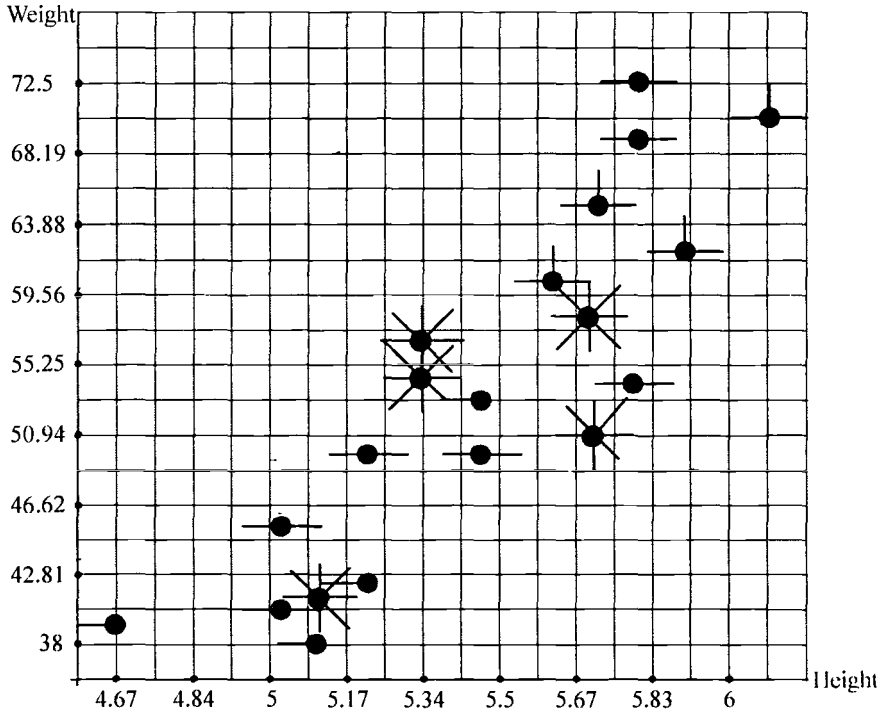


Fig. 99.1. *A sunflower plot for data in Table 17.1 showing Height, Weight and Age of the Individuals*

99.4. Advantages

(*a*) The graphs can be used to study the relationship between X and Y variables.

(*b*) It can also be used to study the variation between X and Y in presence of a third variable which may be either categorical or numerical.

(*c*) It is helpful in detection of outliers.

(*d*) It can be used for the visualisation of three numerical variables or two numerical variables and one categorical variable in two dimensions.

99.5. Disadvantages

(*a*) As the number of observations increases the sunflower plot may lead to over-plotting.

(*b*) Since the magnitude of the third variable is represented by the number of petals so for numerical variables the representation will be only approximate.

99.5. Related Techniques

(*a*) Glyph Plot;

(*b*) Bubble Plot; and

(*c*) Three Dimensional Scatter Plot.

100. SUNFLOWER PLOT (CATEGORICAL)

100.1. Definition and Description

The categorical sunflower plot is same as that of the sunflower plot the only difference is that here an additional variable can be represented in the plot. But the additional variable should be a categorical variable. The sunflower plot is drawn first in the manner discussed in the earlier section and then the sunflowers are colored differently based on the value of the categorical variable. Here different colors are used to represent the different categories, and accordingly we get sunflowers of different colors in the plot. Thus this plot can be used to represent four variables two of which are numerical variable and the third one is either a numerical variable or a categorical variable and the fourth one is a categorical variable.

100.2. Working Data

The data used for drawing the plot is same as that used to draw the bubble plot that is given in Table 18.1. The data comprises of heights, weights, age and sex of 30 individuals collected non-randomly from some known individuals.

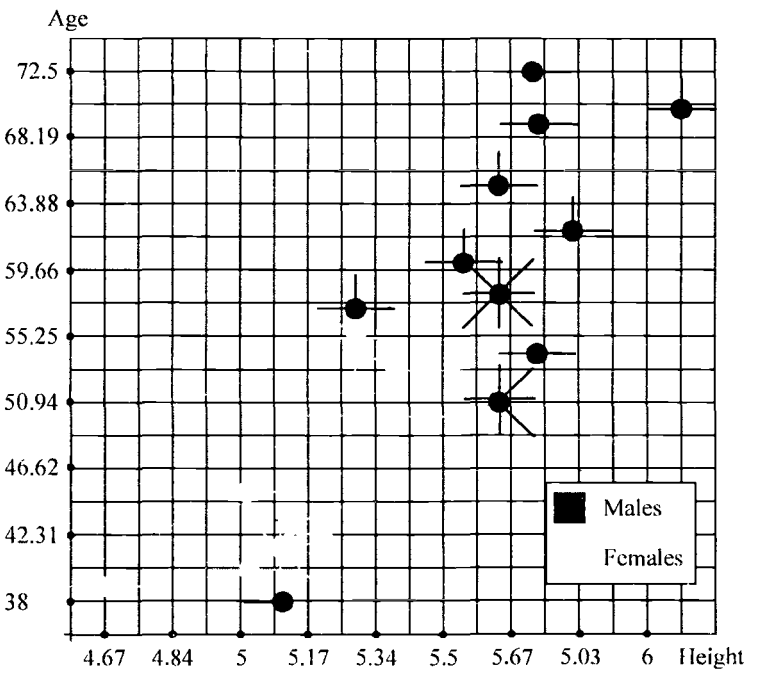


Fig. 100.1. *A categorical sunflower plot for data in Table 18.1*

100.3. Axes

X Axis: It can represent the values of the variable that we suspect may have a relation to the response variable. Here the height is represented along the X axis.

Y Axis: The vertical axis consists of that variable which we consider as the response variable. Here weight is represented along Y axis.

The third variable is represented by whiskers (or rays) that come out from each of the plotted symbols. The number of whiskers that comes out from each symbol is proportional to the value of the third variable i.e. age in this case. A maximum of eight petals are allowed to come out from a sunflower. The more the age more is the number of petals from each sunflower.

The fourth variable is a categorical variable and is represented by the color of the sunflower. Here a black sunflower is used to represent the Males and the grey sunflower is used to represent Females

100.4. Advantages

(*a*) The graphs can be used to study the relationship between X and Y variables.

(*b*) It can also be used to study the variation between X and Y in presence of a two other variables of which one may be either categorical or numerical and the other is a categorical variable.

(*c*) It is helpful in detection of outliers.

(*d*) It can be used for the visualization of three numerical variables and one categorical variable or two numerical variables and two categorical variables in two dimensions.

100.5. Disadvantages

(*a*) As the number of observations increases the sunflower plot may lead to over-plotting.

(*b*) Since the magnitude of the third variable is represented by the number of petals so for numerical variables the representation will be only approximate.

100.6. Related Techniques

(*a*) Categorical Glyph Plot;

(*b*) Categorical Bubble Plot; and

(*c*) Categorical Three Dimensional Scatter Plot.

101. SUNRAY ICON PLOT

101.1. Definition and Description

Sunray icon plots are icon plots used to display multivariate observations. Here corresponding to each multivariate observation we draw a central point with some rays coming out from it. Here, each observation variable is represented by a ray that comes out form a central point. The central point is sometimes expanded to form a small circle. Thus the icon corresponding to each observation looks like a sun. So the icon can be termed as sunray icon plot. These rays are equally spaced around the central point. The angle between any two adjacent rays would be $360^0/p$, where p is the number of variable in the system. The length of the rays is proportional to the value of the variable it represents. Several radii are then equally spaced in such a way that after extending the rays from the radii there is no possibility of over-plotting. The different rays of the sun may be differently colored for each variable. Such use of color makes it easier to identify a particular variable in a particular icon. Here the rays which are used for the representation of the values of the different variable are drawn in a comparative scale.

101.2. Working Data

The data used for this purpose is provided in Table 22.1. The data set is taken from Weber (1973) where the protein consumption in 25 European countries for nine food groups is given.

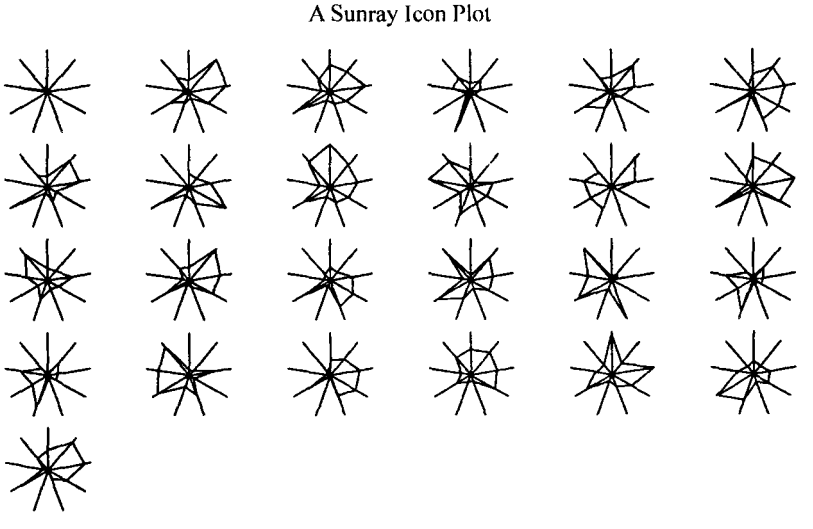


Fig. 101.1. *The sunray icon plot for the protein consumption data*

101.3. Axes

In this plot no axes are used so the order of the variables in a particular sunray is considered either in the clockwise or in the anti-clockwise direction. In this case we have arranged them in clockwise fashion.

101.4. Advantages

(*a*) A sunray icon plot is easier to interpret. If the size of the sun is large then we can understand that the observation has high values of the variable and vise versa.

(*b*) In case of sunray icon plot an additional variable will lead to an additional ray that will come out from the central point and will not occupy any extra space unlike other icon plots like column icon plot or profile plot etc.

(*c*) STATISTICA has the option of drawing such a plot.

101.5. Disadvantages

(*a*) Since in a sunray plot the rays correspond to different variables so each ray will have a different scale of measurement. Thus we cannot compare the performance of the different variables within an observation. To get rid of this problem one may think of standardizing the variables and draw the plot.

(*b*) The variables in a particular observation are difficult to be identified separately if all the rays are of the same color. This can be avoided in case separate color is used for each variable.

101.6. Related Techniques

(*a*) Profile Icon Plot;

(*b*) Star Icon Plot;

(*c*) Chernoff Faces; and

(*d*) Column Icon Plot.

REFERENCES

[1] **Agresti, A. (1989).** Tutorial on modeling ordered categorical response data. *Psychological Bulletin,* 105, 290-301.

[2] **Altman, D.G. (1991).** *Practical Statistics for Medical Research,* London: Chapman and Hall, 286.

[3] **Anderson, T. W. (1958).** *An Introduction to Multivariate Statistical Analysis.* New York: Wiley, 58.

[4] **Andrews, D. F. (1972).** Plots of high dimensional data. *Biometrics,* 28, 125-136.

[5] **Barlett , M.S. (1936).** Some notes on insecticide tests in the laboratory and in the field. *Journal of the Royal Statistical Society, Supplement,* 3, 185-194.

[6] **Behrens, J. and Yu, C. H. (1995).** Visualization techniques of different dimensions. http://seamonkey.ed.asu.edu/~behrens/asu/reports/compre/comp1.html

[7] **Box, G. E. P. and Cox, D. R. (1964).** An Analysis of Transformations. *Journal of the Royal Statistical Society,* 211-243.

[8] **Chernoff, H. (1973).** Using faces to represent points in k-dimensional space graphically. *Journal of American Statistical Association.* 68, 361-368.

[9] **Cleveland, W. S. (1993).** *Visualizing Data.* Hobart Press, Summit, New Jersey.

[10] **Cleveland, W.S. and McGill, R. (1984).** Graphical perception: theory, experimentation and application to the development of graphical methods. *Journal of American Statistical Association,* 79, 531-554.

[11] **Cohen, A. (1980).** On the Graphical Display of the Significant Components in a Two-Way Contingency Table. *Communications in Statistics—Theory and Methods,* A9, 1025-1041.

[12] **David, F. N. (1971).** *A First Course in Statistics,* 2nd Edition, Griffin.

[13] **Doksum, K. (1977).** *Statistica Neerlandica,* 31, 53-68.

[14] **Dubey, S. D. (1966).** Graphical tests for discrete distributions. *The American Statistician,* June, 23-24.

[15] **Dupont, W.D. and Plummer, W.D. (2002).** Density distribution sunflower Plots. *Journal of Statistical Software,* http://www.jstatsoft.org.

[16] **Edwards, D.E. and Kreiner, S. (1983).** The analysis of contingency tables by graphical models. *Biometrika,* 70, 553-565.

[17] **Filliben, J. J. (1997).** *Dataplot Reference Manual.* Statistical Engineering Division, Information Technology Laboratory, NIST.

[18] **Filliben, J. J., Cetinkunt, Yu and Dommenz (1993).** *Explanatory Data Analysis Techniques as Applied to a High-Precision turning Machine,* Elsevier, New York, 199-223.

[19] **Fisher, N.L. and Switzer, P. (1985).** Chi-plots for Assessing Dependence. *Biometrika*, 72, 253-265.

[20] **Freni-Titulauer, L. W. J and Louv, W. C. (1984).** Comparison of some graphical methods for explanatory multivariate data analysis. *The American Statistician*, 38, 184-188.

[21] **Friendly, M. (1994a).** A fourfold display for 2 by 2 by K tables. Technical Report 217, York University, Psychology Dept.

[22] **Friendly, M. (1994b).** SAS/IML graphics for fourfold displays. *Observations*, Vol. 3, No. 4, 47-56

[23] **Friendly, M. (2002).** Corrgrams: Exploratory displays for correlation matrices. *The American Statistician. Vol. 5.*

[24] **Gupta, A. K. (1952).** Estimation of mean and standard deviation of a normal population from a censored sample. *Biometrika*, 39, 260-273.

[25] **Hartigan, J. A. and Kleiner, B. (1981).** Mosaics for contingency tables. In W. F. Eddy (ed.), *Computer Science and Statistics: Proceedings of the 13th Symposium on the Interface*, 268-273. New York, NY: Springer-Verlag.

[26] **Henry, G. T. (1995).** *Graphing Data: Techniques for Display and Analysis.* Applied Social Science Research Method Series: Vol. 36, Thousand Oaks, CA: Sage.

[27] **Hoaglin, D.C. (1980).** A Poissonness plot. *The American Statistician,* Volume 34, 3, 146-149.

[28] **Inselberg, A. (1985).** The plane with parallel coordinates. *The Visual Computer*, vol. 1, 69-91.

[29] **Jacoby, W. (1998).** *Statistical Graphics for Visualizing Multivariate Data.* Sage University Papers Series on Quantitative Applications in the Social Sciences, 07 - 120. Thousand Oaks, CA: Sage.

[30] **Jarrett, R. G. (1979).** A note on the interval between coal mining disasters. *Biometrika,* 66. 191-193.

[31] **Jeffers, J. N. R. (1978).** *An Introduction to System Analysis with Ecological Applications*. London: Edward Arnold.

[32] **Johnson, R. A. and Wichern, D. W. (1992).** *Applied Multivariate Statistical Analysis,* 3rd Edition, Prentice-Hall International, N. J, USA.

[33] **Kemp, A.W. and Kemp, C.D. (1991).** Weldon's dice data revisited. *American Statistician,* 45, 216-222.

[34] **Kotz, S. and Johnson, N. L. (1982).** *Encyclopedia of Statistical Sciences.* Volume 1. John Wiley and Sons, Inc., 85-86.

[35] **Kotz, S., and Johnson, N.L., eds. (1988).** *Encyclopedia of Statistical Sciences*, John Wiley and Sons, New York.

[36] **Mayer, D., Zeileis, A. and Hornik, K. (2003).** Visualizing independence using extended association plot. *Proceedings of the Third International Workshop on Distributed Statistical computing.* March 20-22, 2003, Vienna, Austria.

[37] **McGill, R., Tukey, J. W. and Larsen, W. A. (1978).** Variations of box plots. *The American Statistician,* 32, 12-16.

[38] **Medhi. J. (1994).** *Stochastic Process,* pp. 128, Wiley Eastern Limited, New Delhi, India.

[39] **Ord, J. K. (1967).** Graphical methods for a class of discrete distributions. *Journal of the Royal Statistical Society, Series A,* 130: 232-238.

[40] **Pareto, V. (1897).** Cours d' Economic Politique. Lausanne and Paris: Rouge and Cie.

[41] **Rao, G. V. (1971).** A test for the fitting of some discrete distribution. *Publication de l'Institut de Statistique de l'Universite' de Paris,* 20, 121-128.

[42] **Rao, C. R. (1948).** Test of significance in multivariate analysis. *Biometrika,* 35, 58 - 79.

[43] **Riedwyl, H. and Schiipbach, M. S. (1983).** Graphisehe darstellung von kontingenztafeln. *Technical Report 12,* Institute for Mathematical Statistics, University of Bern, Bern, Switzardland.

[44] **Snedecor, G.W. (1956).** *Statistical Methods.* Iowa State University Press, Ames, Iowa, 1956.

[45] **Spear, M.E. (1952).** *Charting Statistics.* New York: McGraw-Hill.

[46] **Stephens, M.A. (1974).** EDF Statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association,* 69, pp. 730 –737.

[47] **Student (1906).** On the error of counting with a haemocytometer. *Biometrika,* 5, 351-360.

[48] **Tukey, J. W. (1970).** *Explanatory Data Analysis (Limited Preliminary Edition), Vol. 1,* Ch. 5, Reading, Mass: Addison-Wesley Publishing Co.

[49] **Tukey, J. W. (1977).** *Explanatory Data Analysis* (First Edition), Vol. 1, Ch. 5, Reading, Mass: Addison-Wesley Publishing Co.

[50] **Walker, A.M. and Lanes, S.F. (1991).** Misclassification of Covariates. *Statistics in Medicine,* 10, 1181–1196, Table 1.

[51] **Weber, A. (1973).** Agrarpolitik im Spanungsfeld der Internationalon Ernaehrungspolitik, Institut Fuer Agrar Politik und Marktlehre, Kiel.

GLOSSARY

The glossary should not be considered as a list of definitions but are kept in the form of ready reckoner. This is done in order to provide a quick idea to users relatively new in this domain of Statistics. Also some unfamiliar words suddenly come upon in the text may be referred to, for a little more details in this section.

Accessories of Graph	Legends, title, axes labels etc. that help the graphic designer to develop a better presentation.
Abscissa	The x coordinate of a point, *i.e.*, the shortest distance between the point to the y axis.
Attributes	Qualitative variables, those which cannot be measured directly.
Auto Correlation	Correlation of a data set against a shifted version of itself.
Axes titles	Explanatory titles to understand the type of data represented.
Axis	The calibrated sides of a graph.
Breaking of scales	A break must be shown on the vertical scale or in the horizontal scale at a point that will not interfere with the plotting of the data.
Confidence Interval	Range of the parameter determined from sampled data.
Categorical Data	Used in case of qualitative variables. The variable is divided into various mutually exclusive groups.
Contingency	A tabular representation for the frequency of different attributes.
Correlation	A measure of the extent of relationship between two variables.
Curve Legends	In case of more than one curve or bar etc. in a graph they are differentiated by the accompanying legends.
Data Density	The number of entries in the data set divided by the area of data graphic.
Data Labels	Numerical values of the points plotted in the graph at close proximity to the point where it is represented.
Data Source	The reference from which the data is collected.
Dimension	Measure of a physical quantity like, length, breadth etc.

Dimension reduction	Representation of multivariate data in two dimensions.
Dynamic Graphs	Graphs which have movement with time. Generally possible in case of digital graphs.
Expected frequencies	The frequencies that are supposed to be attained by a variable under a given model.
Fitting	Finding a suitable mathematical relationship between two or more variables based on a data set.
Frame	The outer boundary of the graph.
Grid Lines	Some lines drawn parallel to x axis as well as y axis inside the graphed area after regular intervals.
Hyper Dimensional Data	Multi-variate data. Data on more than three variables.
Independent variable	The variable whose values are altered in an experiment.
Kolmogrov – Smirnov test	A non-parametric test to check the goodness of fit of a data to an assumed continuous probability distribution.
Legends	Also called as scale designation, used to indicate the scale of the graph. Usually placed at the right bottom of the graph without overlapping the figures.
Lie Factor	A measure of the distortion in a graph.
Log-log scale	Data converted into logarithms and represented along both the axes.
Ordinate	The y coordinate of a point, *i.e.*, the shortest distance between the point and the x axis.
Outlier	An observation that is unusually large or small compared to the other values of a set of observations.
Over plotting	When the plotted points in a graph falls on one another such that the actual coordinates are difficult to be read from the graph for a large number of points.
Rectilinear coordinate system	The usual coordinate system based on x axis and y axis running perpendicular to each other dividing the plotted area into four quadrants.
Response variable	Dependent variable, a variable whose measurement depends on the value of the other variable.
Semi-log Scale	Data to be represented along the y axis is converted into corresponding logarithm before plotting.

Scale divisions	Used in rectilinear coordinate charts, each of the two axes are marked off in equal units, starting from the origin.
Scale figures	Figures that run from left to right on the horizontal scale and from bottom to top on the vertical scale for the plotting of the values.
Semi logarithmic Charts	An arithmetic scale is considered along the x axis and a logarithmic scale is considered along the Y axis.
Slope	The gradient of a straight line.
Tic Mark	Small mark used to form calibration marks on the axis.
Time Series	Data arranged in chronological order with respect to time.
Time Scale	Axis that is used for representing time as the independent variable.
Title	A short heading of the graph.
Variable	An expression whose value changes.